Supervised Learning on Bakary Data Using WEKA

Waleed Pervaiz
CSE 352
Outline

- Classification Tool: **WEKA**
- **Waikato Environment for Knowledge Analysis** by The University of Waikato.
- Available on the internet at:
  http://www.cs.waikato.ac.nz/~ml/weka/index.html
The Raw data does give us a lot of information. However, in this form most of this information is useless and doesn’t tell us anything.

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
<td>0.02</td>
<td>4318</td>
<td>7</td>
<td>4</td>
<td>0.1</td>
<td>7.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.48</td>
<td>0.04</td>
<td>29649</td>
<td>10</td>
<td>26</td>
<td>201</td>
<td>62.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.00</td>
<td>5726</td>
<td>14</td>
<td>2</td>
<td>7005</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.40</td>
<td>0.01</td>
<td>29379</td>
<td>13</td>
<td>6407</td>
<td>1.7</td>
<td>10.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td>0.03</td>
<td>2158</td>
<td>1919</td>
<td>6112</td>
<td>3</td>
<td>9.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.07</td>
<td>0.00</td>
<td>324570</td>
<td>13715</td>
<td>10647</td>
<td>42</td>
<td>39.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.02</td>
<td>16244</td>
<td>31969</td>
<td>106331</td>
<td>302</td>
<td>116.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.01</td>
<td>0.06</td>
<td>1632</td>
<td>2314</td>
<td>6</td>
<td>6</td>
<td>12.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.09</td>
<td>0.00</td>
<td>791</td>
<td>2477</td>
<td>6</td>
<td>2</td>
<td>5.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00</td>
<td>0.01</td>
<td>13165</td>
<td>12521</td>
<td>5277</td>
<td>22</td>
<td>59.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.84</td>
<td>0.06</td>
<td>3321</td>
<td>25</td>
<td>7</td>
<td>9</td>
<td>9.11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.45</td>
<td>0.06</td>
<td>2888</td>
<td>111</td>
<td>21</td>
<td>51</td>
<td>80.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.23</td>
<td>0.01</td>
<td>209851</td>
<td>23</td>
<td>76</td>
<td>331</td>
<td>121</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.98</td>
<td>0.06</td>
<td>38791</td>
<td>73</td>
<td>27</td>
<td>96</td>
<td>22.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.03</td>
<td>0.08</td>
<td>13769</td>
<td>11</td>
<td>11</td>
<td>5</td>
<td>25.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.65</td>
<td>0.08</td>
<td>275</td>
<td>37</td>
<td>4</td>
<td>0.6</td>
<td>49.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.45</td>
<td>0.03</td>
<td>184</td>
<td>13</td>
<td>7</td>
<td>2</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.65</td>
<td>0.08</td>
<td>652</td>
<td>10</td>
<td>6</td>
<td>288</td>
<td>185</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.37</td>
<td>0.03</td>
<td>314</td>
<td>9</td>
<td>5</td>
<td>18.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.66</td>
<td>0.09</td>
<td>1222</td>
<td>16</td>
<td>3</td>
<td>5</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>K2O</th>
<th>P2O5</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>Zn</td>
<td>Pb</td>
</tr>
<tr>
<td>Cu</td>
<td>As</td>
<td>Cd</td>
</tr>
</tbody>
</table>
Data Preparation

- To prepare data for pre-processing the following steps were taken.
- Any attributes that have missing data (i.e. more than 20%) will be removed.
- The following attributes were thus removed:
  - Pb
  - As
  - Cd
  - Ni
  - Sc
  - Co
  - Li
  - Mo
Data Preparation

- All other attributes that are missing values were filled in with their **averages** (mean).
- Missing values for the following attributes were inserted:
  - TiO$_2$ (Carbonates) – Mean: 0.005 (Inserted at E58 and E62)
  - P$_2$O$_5$ (Carbonates) – Mean: 0.74 (Inserted at L33)
  - S (Carbonates) – Mean: 423 (Inserted at M52, M66, M75)
  - Zn (Carbonates) – Mean: 16 (Inserted at N46, N53, N67)
  - Cu (Carbonates) – Mean: 3 (Inserted at O37, O43, O46, O48, O52, O55, O57, O60, O63, O67)
  - Cr (Galene) – Mean: 9.6 (Inserted at P85, P87)
  - Cr (Spahlerite) – Mean: 3.6 (Inserted at P90)
  - V (Carbonates) – Mean: 5.2 (Inserted at Q43, Q48, Q51, Q52, Q60, Q62, Q66, Q71, Q75)
  - V (Galene) – Mean: 2.5 (Inserted at Q86)
  - V (Spahlerite) – Mean: 9.4 (Inserted at Q90)
Data Preparation

• For easier reading, class values were replaced with simpler values.
• The following values were changed:
  • R. carbonatées changed to C1
  • Pyrite changed to C2
  • Chalcopyrites changed to C3
  • Galène changed to C4
  • Spahlerite changed to C5
  • Sédiments terrigènes changed to C6
Data Preparation

- Using WEKA, we remove any noisy data that may unnecessarily skew our data and results.
Discretization

- With all the missing data filled in, the noisy data eliminated, we discretize the data using the WEKA tool. (3 equal frequency bins)
Discretization

- Values in the bins were then replaced by specific words:
  - Low
  - Medium
  - High

- This helps in understanding data better.

- Decision Tree algorithms will still work with these non-numerical values.
Experiments

- The following experiments will be carried out on our data:
  - **Full Learning**: Construction of decision trees characterizing all classes.
  - **Contrast Learning**: Using all attributes to compare class $C_1$ with the rest of the classes.
  - **Limited Learning**: Construction of decision tree using only the major attributes.
Experiment 1 - Results

- Experiment 1: Full Learning
- Decision Trees were generated using the J48 algorithm.
We got the following discriminant rules.

- IF Ta="Low" AND Lu="Low" AND Fe₂O₃="Low" THEN Class="C₁"
- IF Ta="Low" AND Lu="Low" AND Fe₂O₃="Medium" THEN Class="C₁"
- IF Ta="Low" AND Lu="Low" AND Fe₂O₃="High" AND Cu="Low" THEN Class="C₁"
- IF Ta="Low" AND Lu="Low" AND Fe₂O₃="High" AND Cu="Medium" THEN Class="C₂"
- IF Ta="Low" AND Lu="Low" AND Fe₂O₃="High" AND Cu="High" AND Cr="Low" THEN Class="C₁"
- IF Ta="Low" AND Lu="Low" AND Fe₂O₃="High" AND Cu="High" AND Cr="Medium" THEN Class="C₁"
- IF Ta="Low" AND Lu="Low" AND Fe₂O₃="High" AND Cu="High" AND Cr="High" THEN Class="C₄"
We got the following discriminant rules.
- IF Ta="Low" AND Lu="Medium" THEN Class="C1"
- IF Ta="Low" AND Lu="High" THEN Class="C1"
- IF Ta="Medium" AND CaO="High" THEN Class="C1"
- IF Ta="Medium" AND CaO="Medium" THEN Class="C1"
- IF Ta="Medium" AND CaO="Low" THEN Class="C2"
- IF Ta="High" AND Zn="Low" THEN Class="C1"
- IF Ta="High" AND Zn="Medium" THEN Class="C6"
- IF Ta="High" AND Zn="High" THEN Class="C6"

Predictive Accuracy Determined: 70.58%
Experiment 2 - Results

- Experiment 2: Contrast Learning
- Decision Trees were generated using the J48 algorithm.

```
Classifier output
--------------------
J48 unpruned tree

Fe2O3* = Low: C1 (34.0)
Fe2O3* = Medium: C1 (32.0)
Fe2O3* = High
    CaO = High: C1 (4.0)
    CaO = Medium
        Rb = Medium: NOT C1 (3.0/1.0)
        Rb = Low: C1 (1.0)
        Rb = High: C1 (9.0)
    CaO = Low
        Zn = Low: C1 (2.0)
        Zn = Medium: NOT C1 (8.0)
        Zn = High: NOT C1 (11.0)

Number of Leaves :    9
Size of the tree :    13

Time taken to build model: 0 seconds

--- Evaluation on test split ---
=== Summary ===
Correctly Classified Instances    33    97.0588 %
Incorrectly Classified Instances   1    2.9412 %
```
We got the following discriminant rules.

- IF $\text{Fe}_2\text{O}_3$=“Low” THEN Class=“C1”
- IF $\text{Fe}_2\text{O}_3$=“Medium” THEN Class=“C1”
- IF $\text{Fe}_2\text{O}_3$=“High” AND CaO=“High” THEN Class=“C1”
- IF $\text{Fe}_2\text{O}_3$=“High” AND CaO=“Medium” AND Rb=“Medium” THEN Class=“NOT C1”
- IF $\text{Fe}_2\text{O}_3$=“High” AND CaO=“Medium” AND Rb=“Low” THEN Class=“C1”
- IF $\text{Fe}_2\text{O}_3$=“High” AND CaO=“Medium” AND Rb=“High” THEN Class=“C1”
- IF $\text{Fe}_2\text{O}_3$=“High” AND CaO=“Low” AND Zn=“Low” THEN Class=“C1”
More Discriminant Rules

- We got the following discriminant rules.
  - IF Fe₂O₃="High" AND CaO="Low" AND Zn="Medium" THEN Class="NOT C₁"
  - IF Fe₂O₃="High" AND CaO="Low" AND Zn="High" THEN Class="NOT C₁"

- Predictive Accuracy Determined: 97.06%
Experiment 3 - Results

- **Experiment 3**: Using Major Attributes
- **Decision Trees** were generated using the J48 algorithm.
Discriminant Rules

- We got the following discriminant rules.
  - IF Fe₂O₃=“Low” THEN Class=“C1”
  - IF Fe₂O₃=“Medium” THEN Class=“C1”
  - IF Fe₂O₃=“High” AND CaO=“High” THEN Class=“C1”
  - IF Fe₂O₃=“High” AND CaO=“Medium” THEN Class=“C1”
  - IF Fe₂O₃=“High” AND CaO=“Low” AND Zn=“Low” THEN Class=“C1”
  - IF Fe₂O₃=“High” AND CaO=“Low” AND Zn=“Medium” AND S=“Low” THEN Class=“C6”
  - IF Fe₂O₃=“High” AND CaO=“Low” AND Zn=“Medium” AND S=“Medium” THEN Class=“C6”
More Discriminant Rules

- We got the following discriminant rules.
  - IF Fe$_2$O$_3$="High" AND CaO="Low" AND Zn="Medium" AND S="High" THEN Class="C2"
  - IF Fe$_2$O$_3$="High" AND CaO="Low" AND Zn="High" THEN Class="C6"

- Predictive Accuracy Determined: 82.35%
Discretization for Dataset 2

- With all the missing data filled in, the noisy data eliminated, we use another method of data discretization. (4 equal width bins)
The following experiments will be carried out on our data:

- **Full Learning:** Construction of decision trees characterizing all classes.
- **Contrast Learning:** Using all attributes to compare class C1 with the rest of the classes.
- **Limited Learning:** Construction of decision tree using only the major attributes.
Experiment 1 Dataset 2 - Results

- Experiment 1: Full Learning
- Decision Trees were generated using the J48 algorithm.

```
J48 unpruned tree
-----------------
Ta = Low
  | Lu = Low
    | Fe2O3 = Low: C1 (29.0)
    | Fe2O3 = Medium: C1 (23.0)
    | Fe2O3 = High
    |   | Cu = Low: C1 (1.0)
    |   | Cu = Medium: C2 (2.0)
    |   | Cu = High
    |   |   | Cr = Low: C1 (3.0/1.0)
    |   |   | Cr = Medium: C1 (3.0/1.0)
    |   |   | Cr = High: C4 (3.0)
    | Lu = Medium: C1 (4.0)
    | Lu = High: C1 (4.0/2.0)
Ta = Medium
  | CaO = High: C1 (10.0)
  | CaO = Medium: C1 (3.0/1.0)
  | CaO = Low: C2 (4.0/2.0)
Ta = High
  | Zn = Low: C1 (1.0)
  | Zn = Medium: C6 (3.0)
  | Zn = High: C6 (5.0)

Number of Leaves : 15
Size of the tree : 22
```
Discriminant Rules

- We got the following discriminant rules.
  - IF Ta="Low" AND Lu="Low" AND Fe\textsubscript{2}O\textsubscript{3}="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe\textsubscript{2}O\textsubscript{3}="Medium" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe\textsubscript{2}O\textsubscript{3}="High" AND Cu="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe\textsubscript{2}O\textsubscript{3}="High" AND Cu="Medium" THEN Class="C2"
  - IF Ta="Low" AND Lu="Low" AND Fe\textsubscript{2}O\textsubscript{3}="High" AND Cu="High" AND Cr="Low" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe\textsubscript{2}O\textsubscript{3}="High" AND Cu="High" AND Cr="Medium" THEN Class="C1"
  - IF Ta="Low" AND Lu="Low" AND Fe\textsubscript{2}O\textsubscript{3}="High" AND Cu="High" AND Cr="High" THEN Class="C4"
We got the following discriminant rules.

- IF Ta="Low" AND Lu="Medium" THEN Class="C1"
- IF Ta="Low" AND Lu="High" THEN Class="C1"
- IF Ta="Medium" AND CaO="High" THEN Class="C1"
- IF Ta="Medium" AND CaO="Medium" THEN Class="C1"
- IF Ta="Medium" AND CaO="Low" THEN Class="C2"
- IF Ta="High" AND Zn="Low" THEN Class="C1"
- IF Ta="High" AND Zn="Medium" THEN Class="C6"
- IF Ta="High" AND Zn="High" THEN Class="C6"

Predictive Accuracy Determined: 79.59%
Experiment 2 Dataset 2 - Results

- Experiment 2: Contrast Learning
- Decision Trees were generated using the J48 algorithm.
Discriminant Rules

- We got the following discriminant rules.
  - IF Fe$_2$O$_3$=“Low” THEN Class=“C1”
  - IF Fe$_2$O$_3$=“Medium” THEN Class=“C1”
  - IF Fe$_2$O$_3$=“High” AND CaO=“High” THEN Class=“C1”
  - IF Fe$_2$O$_3$=“High” AND CaO=“Medium” AND Rb=“Medium” THEN Class=“NOT C1”
  - IF Fe$_2$O$_3$=“High” AND CaO=“Medium” AND Rb=“Low” THEN Class=“C1”
  - IF Fe$_2$O$_3$=“High” AND CaO=“Medium” AND Rb=“High” THEN Class=“C1”
  - IF Fe$_2$O$_3$=“High” AND CaO=“Low” AND Zn=“Low” THEN Class=“C1”
More Discriminant Rules

- We got the following discriminant rules.
  - IF Fe$_2$O$_3$=“High” AND CaO=“Low” AND Zn=“Medium” AND Tb=“Low” THEN Class=“NOT C1”
  - IF Fe$_2$O$_3$=“High” AND CaO=“Low” AND Zn=“Medium” AND Tb=“Medium” THEN Class=“C1”
  - IF Fe$_2$O$_3$=“High” AND CaO=“Low” AND Zn=“Medium” AND Tb=“High” THEN Class=“NOT C1”
  - IF Fe$_2$O$_3$=“High” AND CaO=“Low” AND Zn=“High” THEN Class=“NOT C1”

- Predictive Accuracy Determined: 89.79%
Experiment 3 Dataset 2 - Results

- Experiment 3: Using Major Attributes
- Decision Trees were generated using the J48 algorithm.

```
Classifier output

J48 unpruned tree
----------------
Fe2O3* = Low: C1 (34.0)
Fe2O3* = Medium: C1 (32.0)
Fe2O3* = High
  | CaO = High: C1 (4.0)
  | CaO = Medium: C1 (7.0/2.0)
  | CaO = Low
  |   | Zn = Low: C1 (2.0)
  |   | Zn = Medium: NOT C1 (8.0/1.0)
  |   | Zn = High: NOT C1 (11.0)

Number of Leaves : 7
Size of the tree : 10

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

| Correctly Classified Instances | 93            | 94.890 %
| Incorrectly Classified Instances | 5            | 5.102 %
| Kappa statistic               | 0.8456       |
| Mean absolute error           | 0.0607       |
| Root mean squared error       | 0.196        |
```
Discriminant Rules

- We got the following discriminant rules.
  - IF Fe$_2$O$_3$="Low" THEN Class="C1"
  - IF Fe$_2$O$_3$="Medium" THEN Class="C1"
  - IF Fe$_2$O$_3$="High" AND CaO="High" THEN Class="C1"
  - IF Fe$_2$O$_3$="High" AND CaO="Medium" THEN Class="C1"
  - IF Fe$_2$O$_3$="High" AND CaO="Low" AND Zn="Low" THEN Class="C1"
  - IF Fe$_2$O$_3$="High" AND CaO="Low" AND Zn="Medium" THEN Class="NOT C1"
  - IF Fe$_2$O$_3$="High" AND CaO="Low" AND Zn="High" THEN Class="NOT C1"
More Discriminant Rules

- Predictive Accuracy Determined: 94.90%
Accuracy Analysis for both Datasets

- Here is a comparison of the accuracy achieved with each Dataset.

<table>
<thead>
<tr>
<th></th>
<th>Dataset #1</th>
<th>Dataset #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>70.58%</td>
<td>79.59%</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>97.06%</td>
<td>89.79%</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>82.35%</td>
<td>94.90%</td>
</tr>
</tbody>
</table>

- Dataset 1 was carried out using 3 bins – equal frequency discretization.

- Dataset 2 was carried out using 4 bins – equal width discretization.
Conclusion & Thoughts

- High accuracy for a particular value can sometimes be misleading since there is a lot of data (77 records) for C1 as compared to data (21 records) for other classes.

- WEKA produces different rules depending on the techniques used for data preparation.

- Dataset 2 generally had better accuracy. Thus, we can conclude that the 4-bin equal width method was slightly more accurate than the 3-bin equal frequency method.

- Comparing classes with each other gave the best overall accuracy. (i.e. comparing C1 with all other classes)