Supervised Learning with WEKA on Bakary Data

Jason Wu
CSE 352: Artificial Intelligence
Professor Anita Wasilewska
Outline

• Classification Tool: WEKA
• Data Preprocessing
  – Missing Data
  – Discretization
• Experiments
  – Methods
  – Decision Tree
  – Discriminant Rules
  – Predictive Accuracy
• Analysis
Classification Tool: WEKA

- The classification tool that I used was WEKA (the Waikato Environment for Knowledge Analysis)
- It can be obtained from http://www.cs.waikato.ac.nz/~ml/weka/index.html
Data Preprocessing: Missing Data

• Any attributes that had $\geq 80\%$ of its values missing were removed

• Attribute CO had 85% of its values missing
• Attribute Mo had 89% of its values missing

• So, these attributes were removed
Data Preprocessing: Missing Data

- For any attribute that had < 80% of its values missing, the missing values were replaced with the mean (average) value
- These attributes included
  - TiO2, P2O5, S, Zn, Pb, Cu, As, Cd, Cr, Ni, Sc, V, Li
- Example
  - Mean value of TiO2 was 0.043
  - Any missing values in TiO2 were replaced with 0.043
Data Preprocessing: Discretization (using the tool WEKA)

• Data #1:
  – Binning Method
  – 3 Bins
  – Equal Depth (Frequency) Bins

• Data #2:
  – Binning Method
  – 3 Bins
  – Equal Width Bins
Data Preprocessing: Discretization

• For each bin, the values in the bin were replaced with the bin interval

• Example with CaO+MgO
  – Values were separated into 3 bins:
    • (48.64-inf)
    • (45.715-48.64]
    • (-inf-45.715]
  – Values in the “(48.64-inf)” bin were replaced with “(48.64-inf)”
Data Preprocessing

- Class values were replaced with simpler values to read
  - C1 <= R. Carbonatees and R.Carbonatees impures
  - C2 <= Pyrite
  - C3 <= Chalcopyrites
  - C4 <= Galene
  - C5 <= Spahlerite
  - C6 <= Sediments Terrigenes
Experiments

• Experiment 1 (Full Learning)
  – Use all attributes to classify all classes (C1-C6)

• Experiment 2 (Contrast Learning)
  – Use all attributes to contrast class C1 from the others
  – Class values C2-C6 were replaced with NOTC1

• Experiment 3
  – Use only the important attributes to classify all the classes (C1-C6)
  – According to the expert, the important attributes are S, Zn, Pb, Cu, CaO+MgO, CaO, MgO, Fe2O3
Experiment Methods

- Decision trees (and corresponding discriminant rules) were generated with the C4.5 algorithm

- Predictive accuracies were obtained by applying Leave-one-out on the data (and with the final predictive accuracy obtained by taking the average of all runs)
Experiment 1:
Data #1 - Decision Tree
Experiment 1:
Data #1 – 11 Discriminant Rules (1)

• IF Cd = “(8.427586-10.727586]”
  – THEN class = “C1”
• IF Cd = “(-inf-8.427586]” AND Ta = “(-inf-0.05]”
  – AND As = “(28.805741-35.755741]” AND Li = “(-inf-3.2]”
  – THEN class = “C3”
• IF Cd = “(-inf-8.427586]” AND Ta = “(-inf-0.05]”
  – AND As = “(28.805741-35.755741]” AND Li = “(3.2-6.254167]”
  – THEN class = “C1”
• IF Cd = “(-inf-8.427586]” AND Ta = “(-inf-0.05]”
  – AND As = “(28.805741-35.755741]” AND Li = “(6.254167-inf)”
  – THEN class = “C1”
• IF Cd = “(-inf-8.427586]” AND Ta = “(-inf-0.05]”
  – AND As = “(-inf-28.805741]”
  – THEN class = “C2”
Experiment 1:
Data #1 – 11 Discriminant Rules (2)

- IF Cd = “(-inf-8.427586]” AND Ta = “(-inf-0.05]”
  - AND As = “(35.755741-inf)” THEN class = “C1”
- IF Cd = “(-inf-8.427586]” AND Ta = “(0.05-0.15]”
  - AND CaO = “(28.475-inf)” THEN class = “C1”
- IF Cd = “(-inf-8.427586]” AND Ta = “(0.05-0.15]”
  - AND CaO = “(26.87-28.475]” THEN class = “C5”
- IF Cd = “(-inf-8.427586]” AND Ta = “(0.05-0.15]”
  - AND CaO = “(-inf-26.87]” THEN class = “C2”
- IF Cd = “(-inf-8.427586]” AND Ta = “(0.15-inf)”
  - THEN class = “C6”
- IF Cd = “(10.727586-inf)” THEN class = “C4”
Experiment 1:
Data #1 – Predictive Accuracy

• Applying Leave-one-out, the predictive accuracy of these rules was determined to be 88.7755%.
Experiment 1: Data #2 - Decision Tree
Experiment 1:
Data #2 – 6 Discriminant Rules

- IF Nb = “(-inf-3.9]” AND Cu = “(-inf-2335]”
  - AND Zn = “(-inf-10657]” THEN class = “C1”
- IF Nb = “(-inf-3.9]” AND Cu = “(-inf-2335]”
  - AND Zn = “(10657-21313]” THEN class = “C4”
- IF Nb = “(-inf-3.9]” AND Cu = “(-inf-2335]”
  - AND Zn = “(21313-inf)” THEN class = “C4”
- IF Nb = “(-inf-3.9]” AND Cu = “(4670-inf)”
  - THEN class = “C3”
- IF Nb = “(3.9-7.7]” THEN class = “C6”
- IF Nb = “(7.7-inf)” THEN class = “C6”
Experiment 1: Data #2 – Predictive Accuracy

• Applying Leave-one-out, the predictive accuracy of these rules was determined to be 84.6939%.
Experiment 2: 
Data #1 - Decision Tree
Experiment 2:
Data #1 – 9 Discriminantant Rules (1)

- IF Cd = “(8.427586-10.727586]” THEN class = “C1”
- IF Cd = “(-inf-8.427586]” AND Fe2O3* = “(-inf-0.145]”
  - THEN class = “C1”
- IF Cd = “(-inf-8.427586]” AND Fe2O3* = “(0.145-0.335]”
  - THEN class = “C1”
- IF Cd = “(-inf-8.427586]” AND Fe2O3* = “(0.335-inf)”
  - AND CaO = “(28.475-inf)” THEN class = “C1”
- IF Cd = “(-inf-8.427586]” AND Fe2O3* = “(0.335-inf)”
  - AND CaO = “(26.87-28.475]” AND Rb = “(0.225-1.85]”
  - THEN class = “NOTC1”
Experiment 2:
Data #1 – 9 Discriminant Rules (2)

- IF \( Cd = \langle -\infty,-8.427586 \rangle \) AND \( Fe_2O_3^* = \langle 0.335,\infty \rangle \)
  - AND \( CaO = \langle 26.87,28.475 \rangle \) AND \( Rb = \langle -\infty,-0.225 \rangle \)
  - THEN class = “C1”
- IF \( Cd = \langle -\infty,-8.427586 \rangle \) AND \( Fe_2O_3^* = \langle 0.335,\infty \rangle \)
  - AND \( CaO = \langle 26.87,28.475 \rangle \) AND \( Rb = \langle 1.85,\infty \rangle \)
  - THEN class = “C1”
- IF \( Cd = \langle -\infty,-8.427586 \rangle \) AND \( Fe_2O_3^* = \langle 0.335,\infty \rangle \)
  - AND \( CaO = \langle -\infty,26.87 \rangle \) THEN class = “NOTC1”
- IF \( Cd = \langle 10.727586,\infty \rangle \) THEN class = “NOTC1”
Experiment 2:
Data #1 – Predictive Accuracy

• Applying Leave-one-out, the predictive accuracy of these rules was determined to be 92.8571%.
Experiment 2: Data #2 - Decision Tree
Experiment 2:
Data #2 – 6 Discriminant Rules

- IF Nb = “[−∞−3.9]” AND Cu = “[−∞−2335]”
  - AND Zn = “[−∞−10657]” THEN class = “C1”
- IF Nb = “[−∞−3.9]” AND Cu = “[−∞−2335]”
  - AND Zn = “[10657−21313]” THEN class = “NOTC1”
- IF Nb = “[−∞−3.9]” AND Cu = “[−∞−2335]”
  - AND Zn = “[21313−∞)” THEN class = “NOTC1”
- IF Nb = “[−∞−3.9]” AND Cu = “(4670−∞)”
  - THEN class = “NOTC1”
- IF Nb = “[3.9−7.7]” THEN class = “NOTC1”
- IF Nb = “[7.7−∞)” THEN class = “NOTC1”
Experiment 2:  
Data #2 – Predictive Accuracy

- Applying Leave-one-out, the predictive accuracy of these rules was determined to be 84.6939%. 

Experiment 3:
Data #1 - Decision Tree
Experiment 3:
Data #1 – 11 Discriminant Rules (1)

• IF Fe2O3* = “(-inf-0.145]” THEN class = “C1”
• IF Fe2O3* = “(0.145-0.335]” THEN class = “C1”
• IF Fe2O3* = “(0.335-inf)” AND CaO = “(28.475-inf)”
  – THEN class = “C1”
• IF Fe2O3* = “(0.335-inf)” AND CaO = “(26.87-28.475]”
  – THEN class = “C1”
• IF Fe2O3* = “(0.335-inf)” AND CaO = “(-inf-26.87]”
  – AND Zn = “(-inf-4.5]” THEN class = “C1”
• IF Fe2O3* = “(0.335-inf)” AND CaO = “(-inf-26.87]”
  – AND Zn = “(4.5-13.5]” AND Pb = “(1521.75-inf)”
  – THEN class = “C2”
Experiment 3:
Data #1 – 11 Discriminant Rules (2)

- IF $Fe_{2}O_{3}*$ = “(0.335-inf)” AND $CaO$ = “(-inf-26.87]”
  - AND $Zn$ = “(4.5-13.5]” AND $Pb$ = “(-inf-23.5]”
  - THEN class = “C6”
- IF $Fe_{2}O_{3}*$ = “(0.335-inf)” AND $CaO$ = “(-inf-26.87]”
  - AND $Zn$ = “(4.5-13.5]” AND $Pb$ = “(23.5-1521.75]”
  - THEN class = “C2”
- IF $Fe_{2}O_{3}*$ = “(0.335-inf)” AND $CaO$ = “(-inf-26.87]”
  - AND $Zn$ = “(13.5-inf)” AND $Pb$ = “(1521.75-inf)”
  - THEN class = “C4”
- IF $Fe_{2}O_{3}*$ = “(0.335-inf)” AND $CaO$ = “(-inf-26.87]”
  - AND $Zn$ = “(13.5-inf)” AND $Pb$ = “(-inf-23.5]”
  - THEN class = “C6”
- IF $Fe_{2}O_{3}*$ = “(0.335-inf)” AND $CaO$ = “(-inf-26.87]”
  - AND $Zn$ = “(13.5-inf)” AND $Pb$ = “(23.5-1521.75]”
  - THEN class = “C6”
Experiment 3:
Data #1 – Predictive Accuracy

• Applying Leave-one-out, the predictive accuracy of these rules was determined to be 83.6735%.
Experiment 3:  
Data #2 - Decision Tree
Experiment 3:
Data #2 – 6 Discriminantant Rules

- IF CaO+MgO = \( (39.193333-\infty) \) THEN class = “C1”
- If CaO+MgO = \( (19.836667-39.193333] \)
  - AND Cu = \( (-\infty-2335] \) THEN class = “C1”
- If CaO+MgO = \( (19.836667-39.193333] \)
  - AND Cu = \( (4670-\infty) \) THEN class = “C3”
- IF CaO+MgO = \( (-\infty-19.836667] \)
  - AND Fe2O3* = \( (-\infty-13.8] \) THEN class = “C6”
- IF CaO+MgO = \( (-\infty-19.836667] \)
  - AND Fe2O3* = \( (13.8-27.57] \) THEN class = “C6”
- IF CaO+MgO = \( (-\infty-19.836667] \)
  - AND Fe2O3* = \( (27.57-\infty) \) THEN class = “C4”
Experiment 3:
Data #2 – Predictive Accuracy

• Applying Leave-one-out, the predictive accuracy of these rules was determined to be 85.7143%.
Experiments: Summary of Predictive Accuracy (Leave-one-out)

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Data #1 3 Bins Equal Depth</th>
<th>Data #2 3 Bins Equal Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>88.7755%</td>
<td>84.6939%</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>92.8571%</td>
<td>84.6939%</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>83.6735%</td>
<td>85.7143%</td>
</tr>
</tbody>
</table>
Analysis

• The high rule accuracy can be misleading
  – A lot of data about one class, but not the others
    • 77 records about class C1
    • Only 21 records about classes C2-C6
• Contrasting one class against all others may generate more accurate rules than comparing all classes simultaneously
• Notice how we started out with 48 attributes about each data record
  – In the end, we reduced it to only 3 to 5 attributes to classify the data