MDL-Based Unsupervised Attribute Ranking

Zdravko Markov
Computer Science Department
Central Connecticut State University
New Britain, CT 06050, USA
http://www.cs.ccsu.edu/~markov/
markovz@ccsu.edu
MDL-Based Unsupervised Attribute Ranking

• Introduction (Attribute Selection)
• MDL-based Clustering Model Evaluation
• Illustrative Example (“play tennis” data)
• Attribute Ranking Algorithm
• Hierarchical Clustering Algorithm
• Experimental Evaluation
• Conclusion
Attribute Selection

- Supervised / Unsupervised. Find the smallest set of attributes that
  - maximizes predictive accuracy
  - best uncovers interesting natural groupings (clusters) in data according to the chosen criterion

- Subset Selection / Ranking (Weighting)
  - Computationally expensive: $2^m$ attribute sets for $m$ attributes
  - Assumes that attributes are independent
Supervised Attribute Selection

- **Wrapper methods** create prediction models and use the predictive accuracy of these models to measure the attribute relevance to the classification task.
- **Filter methods** directly measure the ability of the attributes to determine the class labels using statistical correlation, information metrics, probabilistic or other methods.
- **There exist numerous methods** in this setting due to the wide availability of model evaluation criteria in supervised learning.
Unsupervised Attribute Selection

- **Wrapper methods** evaluate a subset of attributes by the quality of clustering obtained by using these attributes.

- **Filter methods** explore classical statistical methods for dimensionality reduction, like PCA and maximum variance, information-based or entropy measures.

- There exist very few methods in this setting generally because of the difficulty to evaluate clustering models.
Clustering Model Evaluation

Chapter 4: Evaluating Clustering
- MDL-Based Model and Feature Evaluation

http://www.cs.ccsu.edu/~markov/
http://www.cs.ccsu.edu/~markov/dmwdata.zip
http://www.cs.ccsu.edu/~markov/DMWsoftware.zip
Clustering Model Evaluation

• Consider each possible clustering as a hypothesis $H$ that describes (explains) data $D$ in terms of frequent patterns (regularities).

• Compute the description length of the data $L(D)$, the hypothesis $L(H)$, and data given the hypothesis $L(D|H)$.

• $L(H)$ and $L(D)$ are the minimum number of bits needed to encode (or communicate) $H$ and $D$ respectively.

• $L(D|H)$ represents the number of bits needed to encode $D$ if we know $H$.

• If we know the pattern of $H$, no need to encode all its occurrences in $D$, rather we may encode only the pattern itself and the differences that identify each individual instance in $D$.
Minimum Description Length (MDL) and Information Compression

- The more regularity in $D$ the shorter description length $L(D|H)$.
- Need to balance $L(D|H)$ with $L(H)$, because the latter depends on the complexity of the pattern. Thus **the best hypothesis** should
  - minimize the sum $L(H)+L(D|H)$ (**MDL principle**)
  - or maximize $L(D) - L(H) - L(D|H)$ (**Information Compression**)
Encoding MDL

- Hypotheses and data are uniformly distributed and the probability of occurrence of an item out of \( n \) alternatives is \( 1/n \).
- Minimum code length of the message that a particular item has occurred is \(-\log_2 1/n = \log_2 n\) bits.
- The number of bits needed to encode the choice of \( k \) items out of \( n \) possible items is

\[
-\log_2 \frac{1}{\binom{n}{k}} = \log_2 \binom{n}{k}
\]
Encoding MDL (attribute-value)

• Data $D$, instance $X \in D$, $X$ is a set of $m$ attribute values, $|X| = m$
• $T = \bigcup_{X \in D} X$ - set of all attribute values in $D$, $k = |T|$
• Cluster $C_i$ is defined by the set of all attribute values $T_i \subseteq T$ that occur in its members, $C_i = \{ X \in C_i, X \subseteq T_i \}$
• Clustering $H = \{ C_1, C_2, \ldots, C_n \}$ is defined by $\{ T_1, T_2, \ldots, T_n \}$, $k_i = |T_i|$

$$ L(C_i) = \log_2 \left( \frac{k}{k_i} \right) + \log_2 n $$

$$ L(H) = \sum_{i=1}^{n} L(C_i) $$

$$ L(D_i | C_i) = |C_i| \times \log_2 \left( \frac{k_i}{m} \right) $$

$$ L(D | H) = \sum_{i=1}^{n} L(D_i | C_i) $$

$$ \text{MDL}(C_i) = \log_2 \left( \frac{k}{k_i} \right) + \log_2 n + |C_i| \times \log_2 \left( \frac{k_i}{m} \right) $$

$$ \text{MDL}(H) = \sum_{i=1}^{n} \text{MDL}(C_i) $$
# Play Tennis Data

<table>
<thead>
<tr>
<th>ID</th>
<th>outlook</th>
<th>temp</th>
<th>humidity</th>
<th>windy</th>
<th>play</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>2</td>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>3</td>
<td>overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>rainy</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>rainy</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>rainy</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>7</td>
<td>overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>9</td>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>rainy</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>sunny</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>13</td>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>rainy</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>no</td>
</tr>
</tbody>
</table>

\[C_1 = \{1, 2, 3, 4, 8, 12, 14\} \quad \text{(humidity=high)}\]

\[C_2 = \{5, 6, 7, 9, 10, 11, 13\} \quad \text{(humidity=normal)}\]

\[T_1 = \{\text{outlook=sunny, outlook=overcast, outlook=rainy, temp=hot, temp=mild, humidity=high, windy=false, windy=true}\}\]

\[T_2 = \{\text{outlook=sunny, outlook=overcast, outlook=rainy, temp=hot, temp=mild, temp=cool, humidity=normal, windy=false, windy=true}\}.\]
Clustering Play Tennis Data

\[ MDL(C_i) = \log_2 \left( \frac{k}{k_i} \right) + \log_2 n + |C_i| \times \log_2 \left( \frac{k_i}{m} \right) \]

\[ k_1 = |T_1| = 8, \quad k_2 = |T_2| = 9, \quad k = 10, \quad m = 4, \quad n = 2 \]

\[ MDL(C_1) = \log_2 \left( \frac{10}{8} \right) + \log_2 2 + 7 \times \log_2 \left( \frac{8}{4} \right) = 49.39 \]

\[ MDL(C_2) = \log_2 \left( \frac{10}{9} \right) + \log_2 2 + 7 \times \log_2 \left( \frac{9}{4} \right) = 53.16 \]

\[ MDL(\{C_1, C_2\}) = MDL(\text{humidity}) = 102.55 \text{ bits} \]

1. \[ MDL(\text{temp}) = 101.87 \]
2. \[ MDL(\text{humidity}) = 102.56 \]
3. \[ MDL(\text{outlook}) = 103.46 \]
4. \[ MDL(\text{windy}) = 106.33 \]

- Best attribute is \textbf{temp}
MDL Ranker

- Let $A$ have values $v_1, v_2, \ldots, v_p$
- Clustering $\{C_1, C_2, \ldots, C_p\}$, where $C_i = \{X | x_i \in X\}$
- Let $V_i^A = \emptyset$
- For each data instance $X = \{x_1, x_2, \ldots, x_m\}$
  - For each attribute $A$
    - For each value $x_i$
      - $V_i^A = V_i^A \cup \{x_i\}$
      - $k_i = \sum_{j=1}^{m} |V_j^A|$ 
      - Compute $MDL(\{C_1, C_2, \ldots, C_p\})$

- Incremental (no need to store instances)
- Time $O(nm^2)$, $n$ is the number of data instances
- Space $O(pm^2)$, $p$ is the max number of attribute values
- Evaluates 3204 instances with 13195 attributes (trec data) in 3 minutes.
## Experimental Evaluation Data

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Instances</th>
<th>Attributes</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>reuters</td>
<td>1504</td>
<td>2887</td>
<td>13</td>
</tr>
<tr>
<td>reuters-3class</td>
<td>1146</td>
<td>2887</td>
<td>3</td>
</tr>
<tr>
<td>reuters-2class</td>
<td>927</td>
<td>2887</td>
<td>2</td>
</tr>
<tr>
<td>trec</td>
<td>3204</td>
<td>13195</td>
<td>6</td>
</tr>
<tr>
<td>soybean</td>
<td>683</td>
<td>36</td>
<td>19</td>
</tr>
<tr>
<td>soybean-small</td>
<td>47</td>
<td>36</td>
<td>4</td>
</tr>
<tr>
<td>iris</td>
<td>150</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>ionosphere</td>
<td>351</td>
<td>35</td>
<td>2</td>
</tr>
</tbody>
</table>

Java implementations of MDL ranking and clustering available from [http://www.cs.ccsu.edu/~markov/DMWsoftware.zip](http://www.cs.ccsu.edu/~markov/DMWsoftware.zip)
Experimental Evaluation Metrics

- **Average Precision**

  \[
  \text{Average Precision} = \frac{1}{|D_q|} \sum_{k=1}^{|D|} r_k \times \text{PrecisionAtRank}(k)
  \]

- **PrecisionAtRank**

  \[
  \text{PrecisionAtRank}(k) = \frac{1}{k} \sum_{i=1}^{k} r_i \quad r_i = \begin{cases} 
  1 & \text{if } a_i \in D_q \\
  0 & \text{otherwise}
  \end{cases}
  \]

- **Classes-to-clusters accuracy** ("true" cluster membership)

  \begin{align*}
  \text{root} & [5, 9] \\
  \text{temperature=hot} & [2, 2] \\
  \text{outlook=sunny} & [2] \text{ no} \\
  \text{outlook=overcast} & [2] \text{ yes} \\
  \text{temperature=mild} & [4, 2] \\
  \text{windy=FALSE} & [2, 1] \text{ yes} \\
  \text{windy=TRUE} & [2, 1] \text{ yes} \\
  \text{temperature=cool} & [3, 1] \\
  \text{windy=FALSE} & [2] \text{ yes} \\
  \text{windy=TRUE} & [1, 1] \text{ no}
  \end{align*}

  Clusters (leaves): 6
  Correctly classified instances: 11 (78%)
### Average Precision of Attribute Ranking

| Data set         | $|D_q|$ | InfoGain | MDL    | Error | Entropy |
|------------------|-------|----------|--------|-------|---------|
| reuters          | 15    | 0.3183   | 0.1435 | 0.0642| 0.0030  |
| reuters-3class   | 10    | 0.3948   | 0.1852 | 0.1257| 0.0027  |
| reuters-2class   | 7     | 0.5016   | 0.2438 | 0.1788| 0.3073  |
| trec             | 14    | 0.4890   | 0.2144 | 0.0637| 0.0010  |
| soybean          | 16    | 0.6265   | 0.5606 | 0.3871| 0.4152  |
| soybean-small    | 2     | 0.6428   | 0.3500 | 0.0913| 0.1213  |
| iris             | 1     | 1.0000   | 1.0000 | 1.0000| 0.3333  |
| ionosphere       | 9     | 0.6596   | 0.5041 | 0.2575| 0.4252  |

$D_q$ – set of attributes selected by Wrapper Subset Evaluator with Naïve Bayes classifier.  
**InfoGain** – supervised attribute ranking using Information Gain Evaluator.  
**Error** – unsupervised ranking based on evaluating the quality of clustering by the *sum of squared errors*.  
**Entropy** – unsupervised ranking based on the reduction of the entropy in data when the attribute is removed (Dash and Liu 2000).
Classes-To-Clusters Accuracy With Reuters Data

EM

k-means
Classes-To-Clusters Accuracy With Reuters-3class Data

EM

K-means

- MDL ranked
- InfoGain ranked
Classes-To-Clusters Accuracy With Soybean Data

EM

k-means

MDL ranked
InfoGain ranked

EM

k-means
MDL-Based Clustering

Function $MDL-Cluster(D)$

1. Choose attribute $A = \arg \min_i MDL(A_i)$
2. Let $A$ take values $v_1, v_2, \ldots, v_p$
3. Split data $D = \bigcup_{i=1}^n C_i$, $C_i = \{X \mid x_i \in X\}$
4. If $\text{Comp}(A) > \sum_{i=1}^p \text{Comp}(C_i)$ then stop. Return $D$.
5. For each $i = 1, \ldots, n$ Call $MDL-Cluster(C_i)$
Clustering Reuters-2class Data

root (516550.58) [608, 319]
  trade=0 (434956.39) [507, 18]
    rate=0 (266126.68) [339, 18]
      monei=1 (122154.60) [148] money
      monei=0 (161236.34) [191, 18] money
    rate=1 (125589.70) [168]
      currenc=0 (70870.68) [100] money
      currenc=1 (50491.67) [68] money
  trade=1 (204850.37) [301, 101]
    market=0 (157978.80) [186, 39]
      countri=1 (64418.90) [67, 20] trade
      countri=0 (106457.20) [119, 19] trade
    market=1 (106422.43) [115, 62]
      bank=0 (73572.74) [94, 11] trade
      bank=1 (48489.70) [21, 51] money

-----------------------------
Clusters (leaves): 5
Correctly classified instances: 838 (90%)
# Comparing MDL, EM and k-Means

<table>
<thead>
<tr>
<th>Data set</th>
<th>EM</th>
<th></th>
<th>k-Means</th>
<th></th>
<th>MDL-Cluster</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. %</td>
<td>No. of Clusters</td>
<td>Acc. %</td>
<td>No. of Clusters</td>
<td>Acc. %</td>
<td>No. of Clusters</td>
</tr>
<tr>
<td>reuters</td>
<td>43</td>
<td>6</td>
<td>31</td>
<td>13</td>
<td>59</td>
<td>12</td>
</tr>
<tr>
<td>reuters-3class</td>
<td>58</td>
<td>3</td>
<td>48</td>
<td>3</td>
<td>73</td>
<td>7</td>
</tr>
<tr>
<td>reuters-2class</td>
<td>71</td>
<td>2</td>
<td>61</td>
<td>2</td>
<td>90</td>
<td>7</td>
</tr>
<tr>
<td>trec</td>
<td>26</td>
<td>6</td>
<td>29</td>
<td>6</td>
<td>44</td>
<td>11</td>
</tr>
<tr>
<td>soybean</td>
<td>60</td>
<td>19</td>
<td>51</td>
<td>19</td>
<td>51</td>
<td>7</td>
</tr>
<tr>
<td>soybean-small</td>
<td>100</td>
<td>4</td>
<td>91</td>
<td>4</td>
<td>83</td>
<td>4</td>
</tr>
<tr>
<td>iris</td>
<td>95</td>
<td>3</td>
<td>69</td>
<td>3</td>
<td>96</td>
<td>3</td>
</tr>
<tr>
<td>ionosphere</td>
<td>89</td>
<td>2</td>
<td>81</td>
<td>2</td>
<td>80</td>
<td>3</td>
</tr>
</tbody>
</table>
Conclusion

- MDL-ranker without class information performs closely to the InfoGain method, which essentially uses class information.
- Thus, our approach can improve the performance of clustering algorithms in purely unsupervised setting.
- MDL-cluster outperforms EM and k-means on most benchmark data sets.
- Numeric attributes?
- Subset evaluation?
- Non-hierarchical clustering?

Thank You!