Instance Driven Hierarchical Clustering of Document Collections and Classification by Pattern-Based Hierarchical Clustering

Hassan H. Malik and John R. Kender
September 15th, 2008

Computer Science Department
Columbia University in the city of New York
Outline

• Motivation
• Instance-driven Pattern Mining
• IDHC: A More Flexible Pattern-based Hierarchical Clustering Algorithm
• CPHC: Semi-supervised Classification by Pattern-based Hierarchical Clustering
• Conclusions
Outline

• Motivation
• Instance-driven Pattern Mining
• IDHC: A More Flexible Pattern-based Hierarchical Clustering Algorithm
• CPHC: Semi-supervised Classification by Pattern-based Hierarchical Clustering
• Conclusions
Motivation

- Traditional pattern-mining suffers from
  - Frequency-based pattern significance measures
  - Global thresholds

- Pattern-based hierarchical clustering suffers from
  - An unpredictable number of patterns
  - Unnecessary coupling between pattern size and node height
  - Artificial constraints on soft clustering
Motivation - continued

• Inductive classifiers may not fully exploit the distribution of test instances in the context of the whole dataset

• Existing semi-supervised classification algorithm weaknesses
  – Dependence on flat clustering requires the number of clusters to be known in advance
  – Unnecessary step of training a classifier on the expanded training set
Outline

• Motivation
• **Instance-driven Pattern Mining**
• IDHC: A More Flexible Pattern-based Hierarchical Clustering Algorithm
• CPHC: Semi-supervised Classification by Pattern-based Hierarchical Clustering
• Conclusions
Traditional Pattern Mining

• Aims to mine a set of globally significant patterns from the dataset
• Does not consider local pattern significance
• Traditionally uses a frequency-based pattern significance measure (i.e., Support), and a global threshold
• “Closed interesting” itemsets (Malik and Kender ICDM’06) replaced support with an interestingness measure
  – Still, no coverage guarantees
  – Thresholds not as stable on highly correlated datasets
  – An unpredictable number of resulting patterns
Instance-driven Pattern Mining

- Eliminate the global mining step altogether
- Allow each instance to “vote” for its representative size-2 patterns, balancing global and local pattern significance
  - Sort all patterns in decreasing order of local term frequency * global term interestingness
  - Select all patterns with scores exceeding \textit{min\_standard\_deviation}
  - Number of patterns-per-instance upper bounded by a small constant \textit{max\_K}
- Why size-2? Why not size-3 etc.?
Instance-driven Pattern Mining - advantages

- Coverage guaranteed
- No global threshold
- $min_{standard\_deviation}$ robust across datasets (experimented on 16 datasets)
- A small number of highly significant patterns for each instance
  - Central limit theorem for normally distributed scores
  - Chebyshev's inequality for the rest
- Number of size-2 patterns linear to the number of instances
  - $maxK$ provide empirical upper limit guarantee
## Instance-driven Patterns vs. Closed Interesting Itemsets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#instances</th>
<th>#features</th>
<th>Approx. number of size-2 patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>GPHC, MI</td>
</tr>
<tr>
<td>mm</td>
<td>2,521</td>
<td>126,373</td>
<td>2.4 million</td>
</tr>
<tr>
<td>reviews</td>
<td>4,069</td>
<td>126,373</td>
<td>2.6 million</td>
</tr>
<tr>
<td>sports</td>
<td>8,580</td>
<td>126,373</td>
<td>1.4 million</td>
</tr>
<tr>
<td>tr11</td>
<td>414</td>
<td>6,429</td>
<td>4.3 million</td>
</tr>
<tr>
<td>tr12</td>
<td>313</td>
<td>5,804</td>
<td>3.6 million</td>
</tr>
<tr>
<td>tr23</td>
<td>204</td>
<td>5,832</td>
<td>7.6 million</td>
</tr>
<tr>
<td>tr31</td>
<td>927</td>
<td>10,128</td>
<td>7.0 million</td>
</tr>
</tbody>
</table>
Outline

- Motivation
- Instance-driven Pattern Mining
- **IDHC: A More Flexible Pattern-based Hierarchical Clustering Algorithm**
- CPHC: Semi-supervised Classification by Pattern-based Hierarchical Clustering
- Conclusions
Instance-driven Pattern-based Hierarchical Clustering

• Each size-2 pattern forms an initial cluster, patterns added to their selected-pattern-clusters
• Use instance-to-cluster relationships to prune duplicate (in content) clusters
  – Merge labels of duplicate clusters being removed, enhancing the cluster labels
• Generate rest of the cluster hierarchy by iteratively refining clusters
  – Make patterns progressively longer, and cluster memberships progressively sparser
  – Maintain instance-to-cluster pointers for local-only processing
Cluster Refinement – an example

(a) A transaction dataset as running example

<table>
<thead>
<tr>
<th>Instance ID</th>
<th>Features and their local frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>(A:2), (B:4), (D:1), (H:2), (J:4), (L:1)</td>
</tr>
<tr>
<td>T2</td>
<td>(A:3), (C:1), (D:6), (E:1), (G:4)</td>
</tr>
<tr>
<td>T3</td>
<td>(B:2), (C:3), (D:1), (I:5), (K:2)</td>
</tr>
<tr>
<td>T4</td>
<td>(B:3), (C:1), (D:2), (E:4), (J:3), (K:3), (L:2)</td>
</tr>
<tr>
<td>T5</td>
<td>(B:7), (C:2), (D:1), (H:3), (I:2)</td>
</tr>
<tr>
<td>T6</td>
<td>(A:1), (B:1), (C:1), (E:1), (J:3), (K:1)</td>
</tr>
<tr>
<td>T7</td>
<td>(B:9), (C:3), (F:4), (H:5), (J:1), (L:5)</td>
</tr>
<tr>
<td>T8</td>
<td>(C:6), (D:2), (G:1), (I:1), (K:3)</td>
</tr>
<tr>
<td>T9</td>
<td>(B:3), (D:2), (J:4), (K:1), (L:8)</td>
</tr>
<tr>
<td>T10</td>
<td>(B:2), (D:7), (F:3), (I:6)</td>
</tr>
<tr>
<td>T11</td>
<td>(C:1), (E:1), (F:1), (G:2), (H:1), (I:4), (J:1)</td>
</tr>
</tbody>
</table>

(b) Global significance values of some size-2 patterns using Added Value (transformed to positive scale)

<table>
<thead>
<tr>
<th>Pattern</th>
<th>AV</th>
<th>Pattern</th>
<th>AV</th>
<th>Pattern</th>
<th>AV</th>
<th>Pattern</th>
<th>AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(B, D)</td>
<td>0.52</td>
<td>(B, E)</td>
<td>0.38</td>
<td>(B, J)</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(B, K)</td>
<td>0.57</td>
<td>(E, J)</td>
<td>0.70</td>
<td>(J, K)</td>
<td>0.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(E, K)</td>
<td>0.77</td>
<td>(J, L)</td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(D, E)</td>
<td>0.38</td>
<td>(K, L)</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(E, L)</td>
<td>0.38</td>
<td>(D, I)</td>
<td>2.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(F, I)</td>
<td>(C, I)</td>
<td>(H, I)</td>
<td>3.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(c) Instance pattern selection

<table>
<thead>
<tr>
<th>Instance ID</th>
<th>#size-2 patterns</th>
<th>Significance range</th>
<th>Min sig.</th>
<th>Selected patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>15</td>
<td>0.52</td>
<td>2.42</td>
<td>(B, J) (J, L), (H, J)</td>
</tr>
<tr>
<td>T2</td>
<td>10</td>
<td>0.77</td>
<td>2.38</td>
<td>(D, G) (A, D)</td>
</tr>
<tr>
<td>T3</td>
<td>10</td>
<td>0.78</td>
<td>2.29</td>
<td>(C, I)</td>
</tr>
<tr>
<td>T4</td>
<td>21</td>
<td>0.57</td>
<td>2.46</td>
<td>(E, J) (J, L)</td>
</tr>
<tr>
<td>T5</td>
<td>10</td>
<td>0.59</td>
<td>2.61</td>
<td>(B, H) (B, D)</td>
</tr>
<tr>
<td>T6</td>
<td>15</td>
<td>0.33</td>
<td>1.40</td>
<td>(E, J) (B, J) (J, K)</td>
</tr>
<tr>
<td>T7</td>
<td>15</td>
<td>0.90</td>
<td>5.40</td>
<td>(B, L)</td>
</tr>
<tr>
<td>T8</td>
<td>10</td>
<td>0.71</td>
<td>2.70</td>
<td>(C, G) (C, K)</td>
</tr>
<tr>
<td>T9</td>
<td>10</td>
<td>0.85</td>
<td>5.72</td>
<td>(J, L) (B, L)</td>
</tr>
<tr>
<td>T10</td>
<td>10</td>
<td>1.19</td>
<td>3.72</td>
<td>(D, I) (F, I)</td>
</tr>
<tr>
<td>T11</td>
<td>21</td>
<td>0.38</td>
<td>2.13</td>
<td>(G, I) (F, I) (C, I) (H, I)</td>
</tr>
</tbody>
</table>
Cluster Refinement – an example
Instance-driven Hierarchical Clustering - advantages

• Number of initial patterns predictable
• Cluster refinement avoids global processing
• No coupling between node heights and pattern-lengths
  – More meaningful cluster labels
• More flexible soft clustering
  – Instances allowed to exist at multiple levels in the hierarchy
  – Instances not forced to their longest-pattern clusters
• Parameter values robust across datasets
## Clustering Quality on Text Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Fscores</th>
<th></th>
<th>Entropies</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bi-k $I_2$</td>
<td>GPHC</td>
<td>Ours</td>
<td>bi-k $I_2$</td>
</tr>
<tr>
<td>reuters</td>
<td>0.835</td>
<td><strong>0.851</strong></td>
<td>0.846</td>
<td>0.075</td>
</tr>
<tr>
<td>classic</td>
<td>0.782</td>
<td><strong>0.88</strong></td>
<td>0.759</td>
<td>0.06</td>
</tr>
<tr>
<td>hitech</td>
<td>0.528</td>
<td>0.54</td>
<td><strong>0.544</strong></td>
<td>0.224</td>
</tr>
<tr>
<td>k1a</td>
<td>0.668</td>
<td>0.654</td>
<td><strong>0.676</strong></td>
<td>0.106</td>
</tr>
<tr>
<td>k1b</td>
<td>0.882</td>
<td><strong>0.903</strong></td>
<td>0.897</td>
<td>0.042</td>
</tr>
<tr>
<td>la12</td>
<td>0.741</td>
<td>0.661</td>
<td><strong>0.748</strong></td>
<td>0.12</td>
</tr>
<tr>
<td>mm</td>
<td>0.774</td>
<td><strong>0.943</strong></td>
<td>0.909</td>
<td>0.073</td>
</tr>
<tr>
<td>ohscal</td>
<td><strong>0.601</strong></td>
<td>0.53</td>
<td>0.554</td>
<td>0.198</td>
</tr>
<tr>
<td>re0</td>
<td>0.61</td>
<td><strong>0.672</strong></td>
<td>0.615</td>
<td>0.115</td>
</tr>
<tr>
<td>reviews</td>
<td>0.801</td>
<td>0.818</td>
<td><strong>0.833</strong></td>
<td>0.073</td>
</tr>
<tr>
<td>sports</td>
<td>0.882</td>
<td><strong>0.886</strong></td>
<td>0.87</td>
<td>0.03</td>
</tr>
<tr>
<td>tr11</td>
<td><strong>0.795</strong></td>
<td>0.519</td>
<td>0.79</td>
<td>0.107</td>
</tr>
<tr>
<td>tr12</td>
<td>0.689</td>
<td>0.604</td>
<td><strong>0.769</strong></td>
<td>0.133</td>
</tr>
<tr>
<td>tr23</td>
<td>0.667</td>
<td>0.487</td>
<td><strong>0.679</strong></td>
<td>0.136</td>
</tr>
<tr>
<td>tr31</td>
<td>0.837</td>
<td>0.584</td>
<td><strong>0.84</strong></td>
<td>0.041</td>
</tr>
<tr>
<td>wap</td>
<td><strong>0.683</strong></td>
<td>0.663</td>
<td>0.67</td>
<td>0.106</td>
</tr>
<tr>
<td>average</td>
<td>0.736</td>
<td>0.7</td>
<td><strong>0.75</strong></td>
<td>0.102</td>
</tr>
</tbody>
</table>
Outline

• Motivation
• Instance-driven Pattern Mining
• IDHC: A More Flexible Pattern-based Hierarchical Clustering Algorithm
• CPHC: Semi-supervised Classification by Pattern-based Hierarchical Clustering
• Conclusions
Existing Classifiers

- **Inductive classifiers**
  - Use instances in the training set to obtain a classification model
  - Use this classification model to determine class labels for test instances
  - Cons:
    - May not fully exploit the distribution of test instances in the context of the whole dataset
    - Poor classification performance when training data is sparse

- **Semi-supervised classification algorithms**
  - First (flat) cluster training and test sets together
  - Use the resulting clustering solution to enhance the training set
  - Cons:
    - Flat clustering requires the number of clusters to be known in advance
    - Extra step of training a classifier on the expanded training set
Pattern-based Cluster Hierarchies and Significance of Pattern Lengths

- Lower overall Entropy = a higher percentage of nodes that contain most instances that belong to the same ground truth class
- IDHC only assigns instances to their “selected” pattern clusters
  - Intuition: Nodes with longer patterns should have lower Entropies
- Experimented with 4 datasets to understand the class-label distributions over nodes with varying pattern-lengths
Average Node Entropies With Respect to Pattern Sizes

![Graph showing average entropies with respect to pattern sizes.]
The CPHC Classification Algorithm

• **Feature selection**
  – Use a supervised method for training instances
  – Use an unsupervised method for test instances
  – Ensure coverage

• **Clustering**
  – Apply the instance-driven, pattern-based hierarchical clustering algorithm (IDHC) on all training and test instances
  – Track interestingness values

• **Classification**
  – For each test instance $t$, traverse the hierarchy to identify the set $S$ of clusters that contain $t$
  – Use interestingness values of clusters in $S$, and pattern-lengths as weights to compute class scores for $t$
Improving The Chances of Classifying Isolated Test Instances

• Classification model produced by inductive classifiers limited to patterns in training instances
  – No way of classifying isolated test instances

• Improving the chances of classifying such test instances by inducing a type of transitivity
  – Isolated test instances may be clustered together in a “logical” node with test instances that overlap the training set
  – The “logical” node contributes towards score
## Breakeven Performance on Top 10 Reuters 21578 Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Harmony</th>
<th>Find Sim</th>
<th>Naïve Bayes</th>
<th>Bayes Nets</th>
<th>Trees</th>
<th>SVM (linear)</th>
<th>ARC-BC</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>acq</td>
<td>95.3</td>
<td>64.7</td>
<td>87.8</td>
<td>88.3</td>
<td>89.7</td>
<td>93.6</td>
<td>90.9</td>
<td>94.5</td>
</tr>
<tr>
<td>corn</td>
<td>78.2</td>
<td>48.2</td>
<td>65.3</td>
<td>76.4</td>
<td>91.8</td>
<td>90.3</td>
<td>69.6</td>
<td>77.2</td>
</tr>
<tr>
<td>crude</td>
<td>85.7</td>
<td>70.1</td>
<td>79.5</td>
<td>79.6</td>
<td>85</td>
<td>88.9</td>
<td>77.9</td>
<td>90.7</td>
</tr>
<tr>
<td>earn</td>
<td>98.1</td>
<td>92.9</td>
<td>95.9</td>
<td>95.8</td>
<td>97.8</td>
<td>98</td>
<td>92.8</td>
<td>96.5</td>
</tr>
<tr>
<td>grain</td>
<td>91.8</td>
<td>67.5</td>
<td>78.8</td>
<td>81.4</td>
<td>85</td>
<td>94.6</td>
<td>68.8</td>
<td>91.1</td>
</tr>
<tr>
<td>interest</td>
<td>77.3</td>
<td>63.4</td>
<td>64.9</td>
<td>71.3</td>
<td>67.1</td>
<td>77.7</td>
<td>70.5</td>
<td>81</td>
</tr>
<tr>
<td>money-fx</td>
<td>80.5</td>
<td>46.7</td>
<td>56.6</td>
<td>58.8</td>
<td>66.2</td>
<td>74.5</td>
<td>70.5</td>
<td>84.3</td>
</tr>
<tr>
<td>ship</td>
<td>86.9</td>
<td>49.2</td>
<td>85.4</td>
<td>84.4</td>
<td>74.2</td>
<td>85.6</td>
<td>73.6</td>
<td>78.3</td>
</tr>
<tr>
<td>trade</td>
<td>88.4</td>
<td>65.1</td>
<td>63.9</td>
<td>69</td>
<td>72.5</td>
<td>75.9</td>
<td>68</td>
<td>87.9</td>
</tr>
<tr>
<td>wheat</td>
<td>62.8</td>
<td>68.9</td>
<td>69.7</td>
<td>82.7</td>
<td>92.5</td>
<td>91.8</td>
<td>84.8</td>
<td>83.6</td>
</tr>
<tr>
<td>micro-avg</td>
<td>92</td>
<td>64.6</td>
<td>81.5</td>
<td>85</td>
<td>88.4</td>
<td>92</td>
<td>82.1</td>
<td>92.1</td>
</tr>
<tr>
<td>macro-avg</td>
<td>84.5</td>
<td>63.7</td>
<td>74.8</td>
<td>78.8</td>
<td>82.2</td>
<td>87.1</td>
<td>76.7</td>
<td>86.5</td>
</tr>
</tbody>
</table>
### Classification Accuracies on 13 Small and 2 Large UCI Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>FOIL</th>
<th>CPAR</th>
<th>SVM</th>
<th>Harmony</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>anneal</td>
<td>96.9</td>
<td>90.2</td>
<td>83.83</td>
<td>91.51</td>
<td>93.82</td>
</tr>
<tr>
<td>auto</td>
<td>46.1</td>
<td>48</td>
<td>55.5</td>
<td>61</td>
<td>73</td>
</tr>
<tr>
<td>breast</td>
<td>94.4</td>
<td>94.8</td>
<td>96.8</td>
<td>92.42</td>
<td>93.33</td>
</tr>
<tr>
<td>glass</td>
<td>49.3</td>
<td>48</td>
<td>46</td>
<td>49.8</td>
<td>70</td>
</tr>
<tr>
<td>heart</td>
<td>57.4</td>
<td>51.1</td>
<td>60.36</td>
<td>56.46</td>
<td>58.33</td>
</tr>
<tr>
<td>hepatitis</td>
<td>77.5</td>
<td>76.5</td>
<td>81.83</td>
<td>83.16</td>
<td>83.33</td>
</tr>
<tr>
<td>horsecolic</td>
<td>83.5</td>
<td>82.3</td>
<td>83.31</td>
<td>82.53</td>
<td>73.61</td>
</tr>
<tr>
<td>ionoSphere</td>
<td>89.5</td>
<td>92.9</td>
<td>89.44</td>
<td>92.03</td>
<td>92.57</td>
</tr>
<tr>
<td>iris</td>
<td>94</td>
<td>94.7</td>
<td>94.67</td>
<td>93.32</td>
<td>94.67</td>
</tr>
<tr>
<td>pima</td>
<td>73.8</td>
<td>75.6</td>
<td>74.18</td>
<td>72.34</td>
<td>73.16</td>
</tr>
<tr>
<td>tic-tac-toe</td>
<td>96</td>
<td>72.2</td>
<td>70.78</td>
<td>92.29</td>
<td>72.74</td>
</tr>
<tr>
<td>wine</td>
<td>86.4</td>
<td>92.5</td>
<td>94.9</td>
<td>91.94</td>
<td>88.24</td>
</tr>
<tr>
<td>zoo</td>
<td>96</td>
<td>96</td>
<td>86</td>
<td>93</td>
<td>97</td>
</tr>
<tr>
<td><strong>average</strong></td>
<td>80.06</td>
<td>78.06</td>
<td>78.28</td>
<td>80.91</td>
<td>81.83</td>
</tr>
</tbody>
</table>

### Classification Accuracies on 13 Small and 2 Large UCI Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>FOIL</th>
<th>CPAR</th>
<th>SVM</th>
<th>Harmony</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult</td>
<td>82.5</td>
<td>76.7</td>
<td>84.16</td>
<td>81.9</td>
<td>84.95</td>
</tr>
<tr>
<td>mushroom</td>
<td>99.5</td>
<td>98.8</td>
<td>99.67</td>
<td>99.94</td>
<td>99.98</td>
</tr>
<tr>
<td><strong>average</strong></td>
<td>91</td>
<td>87.85</td>
<td>91.92</td>
<td>90.92</td>
<td>92.46</td>
</tr>
</tbody>
</table>
## Classification Accuracies on Sports with Various Parameter Values

<table>
<thead>
<tr>
<th>Harmony (Min support)</th>
<th>75</th>
<th>100</th>
<th>125</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>94.2</td>
<td>94.9</td>
<td>94.3</td>
<td>94.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SVM (C)</th>
<th>2</th>
<th>1</th>
<th>0.5</th>
<th>0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95.79</td>
<td>95.79</td>
<td>95.76</td>
<td>95.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ours (min_supp)</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>96.4</td>
<td>96.24</td>
<td>96.12</td>
<td>95.98</td>
</tr>
</tbody>
</table>
Classification Accuracies on Classic and Re0 with Increasingly Sparser Training Data

![Graphs showing classification accuracies for Classic and Re0 with varying percentage of instances used for training.](Image)
Outline

• Motivation
• Instance-driven Pattern Mining
• IDHC: A More Flexible Pattern-based Hierarchical Clustering Algorithm
• CPHC: Semi-supervised Classification by Pattern-based Hierarchical Clustering
• Conclusions
Conclusions

• Pattern mining
  – Interestingness measures outperform frequency-based measures
  – Instance-driven pattern mining more stable than global pattern mining
    • Local thresholds more robust than global thresholds

• Pattern-based hierarchical clustering
  – Instance-driven approach more stable than global approach
Conclusions - continued

• Pattern-based hierarchical clustering
  – Use instance-to-cluster pointers to avoid global refinement
  – Tight coupling between node heights and pattern lengths unnecessary

• Classification
  – Relying on training data alone may result in suboptimal classification results, specially with sparse training data
Conclusions - continued

• Classification
  – Using a pattern-based cluster hierarchy as a direct mean for semi-supervised classification
    • No need to know the number of clusters in advance
    • No extra step of training on an expanded training set
    • Exploits pattern lengths
    • May improve classification of isolated test instances
Questions?