semi-supervised data organization

Javed Aslam
Sergey Bratus
Virgil Pavlu
motivation: too much data

log analysis:
- lots of data
- hard to search/visualize
- very few labeled records

but
- easy to cluster/classify
- interesting clusters have high density linkage
- lots of similar records
motivation: lots of log data
Filter in Ethereal:
- if we would know what to look for....
- similar records do not necessarily match boolean logic
- filters get too long
- one at a time, loses big picture
  - difficult browsing
TreeView

Snort portscan alerts

- **1339**
  - **1135**
    - dst_port: 445  src_ip: [55]  dst_ip: [75]
  - **70**
  - **26**
  - **22**
    - dst_port: 4899  src_ip: 218.103.195.242  dst_ip: [22]
  - **20**
  - **15**
    - dst_port: 139  src_ip: 129.170.125.243  dst_ip: [8]
  - **15**
    - dst_port: 443  src_ip: 211.5.239.5  dst_ip: [9]
  - **12**
    - dst_port: 1524  src_ip: 192.139.15.34  dst_ip: [12]
  - **9**
    - dst_port: 1  src_ip: 209.15.84.72  dst_ip: [9]
  - **3**
    - dst_port: 8000  src_ip: 194.208.40.120  dst_ip: [2]
  - **3**
    - dst_port: 1080  src_ip: 194.208.40.120  dst_ip: [2]
  - **3**
    - dst_port: 3128  src_ip: 194.208.40.120  dst_ip: [2]
  - **3**
    - dst_port: 8100  src_ip: 194.208.40.120  dst_ip: [2]
  - **3**
    - dst_port: 8080  src_ip: 194.208.40.120  dst_ip: [2]

- **33**
  - src_ip: 194.208.40.120  dst_ip: [2]
  - Apr 15 19:54:10 anon snort: 194.208.40.120 4743 -> 129.170.166.39
  - Apr 15 19:55:00 anon snort: 194.208.40.120 4914 -> 129.170.166.39

Attributes

<table>
<thead>
<tr>
<th>Field</th>
<th>#</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>_day</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>_minute</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>type</td>
<td>SYN</td>
<td></td>
</tr>
<tr>
<td>_month</td>
<td>Apr</td>
<td></td>
</tr>
<tr>
<td>loghost</td>
<td>anon</td>
<td></td>
</tr>
<tr>
<td>_hour</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>_line</td>
<td>Apr 15 19:54:10 anon snort: 194.208.40....</td>
<td></td>
</tr>
<tr>
<td>dst_port</td>
<td>8080</td>
<td></td>
</tr>
<tr>
<td>src_ip</td>
<td>194.208.40.120</td>
<td></td>
</tr>
<tr>
<td>src_port</td>
<td>4743</td>
<td></td>
</tr>
<tr>
<td>flags</td>
<td>******<em>5</em></td>
<td></td>
</tr>
<tr>
<td>program</td>
<td>snort</td>
<td></td>
</tr>
<tr>
<td>dst_ip</td>
<td>129.170.166.39</td>
<td></td>
</tr>
<tr>
<td>_second</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
data organization: trees

- good tree:
  - small branching factor
  - anomalies grouped together
  - branches are different
  - easy browsing
  - able to use feedback
from table to tree
\[ H(X, Y) = H(X) + H(Y) - I(X, Y) = \\
= H(X) + H(Y|X) \]

\[ H(Y|X) = H(X, Y) - H(X) = \\
= H(Y) - I(X, Y) \]

\[ I(X, Y) = H(X, Y) - H(X|Y) - H(Y|X) = \\
= H(X) - H(Y|X) \]
information theory

\[ H(X, Y) = 6.039 \]

- \( H = 6.006 \) for \( src\_ip \)
- \( H = 1.076 \) for \( dst\_port \)
information theory

$H(X,Y) = 6.550$

$H(X,Y) = 9.286$

$H(X,Y) = 6.039$

$H(X,Y) = 9.351$
information theory

H(YI|X)=0.76

H(YI|X)=2.216

H(YI|X)=0.39

H(YI|X)=3.35

Pick me!
Jensen-Shannon divergence

\[ JS(D_1, \ldots, D_n) = H\left( \sum_{k=1}^{n} p_k D_k \right) - \sum_{k=1}^{n} p_k H(D_k) \]

- measures (dis)similarity between several distributions
- almost a distance
- represents information reduction from encoding the distributions together rather than separately
- zero iff all distributions are identical
- better analytical properties than relative entropy
JS divergence for tree

\[ JS_{\pi_i}(\mathcal{R}_1, \ldots, \mathcal{R}_{|F_i|}) = H\left( \sum_{k=1}^{\frac{|F_i|}{2}} p_k \mathcal{R}_k \right) - \sum_{k=1}^{\frac{|F_i|}{2}} p_k H(\mathcal{R}_k) \]

- \( \overline{\mathcal{R}} = \mathcal{R}_1 \cup \ldots \cup \mathcal{R}_{|F_i|} \)
- measures dissimilarities between tree branches
- bar on top of each node indicates the number of records and their class labels (unknown)

\[
\mathcal{R}_k \quad \underbrace{A \quad B}_{p_k}
\]
information bottleneck

- [Tishby, Pereira, Bialek]
  - $X$ is the set of objects to be clustered/compressed
  - $Y =$ relevant feature(s)
  - find cluster $C$ to achieve

$$\arg \min_C I(X; C) - \beta I(Y; C)$$
information bottleneck and JS

- information bottleneck formula

\[
\arg\min_C I(X; C) - \beta I(Y; C)
\]

- applied with JS divergence

\[
F(1) = \arg\min_{\{F_i\mid H(F_i) \neq 0\}} H(F_i) - \beta \cdot JS_{\pi_i}(\mathcal{R}_1, \ldots, \mathcal{R}_{|F_i|}),
\]
semi-supervised

- $L = \text{set of labels provided by the user}$
  - Only a tiny percentage of records will be marked either way.
  - Not all copies of identical records (or very similar) records will be marked

- all quantities of interest become conditionals of $L$:

\[
H(F|L) = H(F) - I(F; L)
\]

\[
JS_{\pi_i}(\mathcal{R}_1, \ldots, \mathcal{R}_{|F_i|}|L) = JS_{\pi_i}(\mathcal{R}_1, \ldots, \mathcal{R}_{|F_i|})
\]

\[
- (I(\overline{\mathcal{R}}; L) - \sum_{k=1}^{|F_i|} p^i_k I(\mathcal{R}_k; L))
\]
semi-supervised

\[ F^{(1)} = \arg \min_{\{F_i | H(F_i) \neq 0\}} H(F_i | L) - \beta \cdot JS_{\pi_i}(R_1, \ldots, R_{|F_i|} | L), \]

\[ \overline{R} = R_1 \cup \cdots \cup R_{|F_i|} \]
results
results
results
## Table 1. Performance with 5% of labeling

<table>
<thead>
<tr>
<th></th>
<th>MDL(ID3)</th>
<th>MDL(JSV)</th>
<th>Err(ID3)</th>
<th>Err(JVS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAL</td>
<td>436</td>
<td>422</td>
<td>0.30</td>
<td>0.28</td>
</tr>
<tr>
<td>BAND</td>
<td>359</td>
<td>343</td>
<td>0.29</td>
<td>0.39</td>
</tr>
<tr>
<td>CAR</td>
<td>707</td>
<td>699</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>CMC</td>
<td>1612</td>
<td>1648</td>
<td>0.68</td>
<td>0.52</td>
</tr>
<tr>
<td>CRX</td>
<td>381</td>
<td>221</td>
<td>0.22</td>
<td>0.14</td>
</tr>
<tr>
<td>MONK</td>
<td>284</td>
<td>284</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>TIC</td>
<td>663</td>
<td>595</td>
<td>0.33</td>
<td>0.25</td>
</tr>
<tr>
<td>VOTE</td>
<td>119</td>
<td>47</td>
<td>0.12</td>
<td>0.04</td>
</tr>
</tbody>
</table>

## Table 2. Performance with 10% of labeling

<table>
<thead>
<tr>
<th></th>
<th>MDL(ID3)</th>
<th>MDL(JSV)</th>
<th>Err(ID3)</th>
<th>Err(JVS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAL</td>
<td>431</td>
<td>422</td>
<td>0.30</td>
<td>0.29</td>
</tr>
<tr>
<td>BAND</td>
<td>346</td>
<td>348</td>
<td>0.28</td>
<td>0.35</td>
</tr>
<tr>
<td>CAR</td>
<td>630</td>
<td>703</td>
<td>0.13</td>
<td>0.18</td>
</tr>
<tr>
<td>CMC</td>
<td>1596</td>
<td>1638</td>
<td>0.60</td>
<td>0.53</td>
</tr>
<tr>
<td>CRX</td>
<td>206</td>
<td>221</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>MONK</td>
<td>284</td>
<td>284</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>TIC</td>
<td>521</td>
<td>589</td>
<td>0.29</td>
<td>0.24</td>
</tr>
<tr>
<td>VOTE</td>
<td>47</td>
<td>47</td>
<td>0.04</td>
<td>0.04</td>
</tr>
</tbody>
</table>
application: ethereal plugin

- kerf.cs.dartmouth.edu
Thank You