Multi-label Classification using Ensembles of Pruned Sets

Jesse Read, Bernhard Pfahringer, Geoff Holmes

University of Waikato
New Zealand

ICDM 2008, December 15, 2008. Pisa, Italy
Introduction

- A set of instances: $D = \{x_0, x_1, \cdots, x_m\}$
- A set of predefined labels: $L = \{l_0, l_1, \cdots, l_n\}$
- Single-label Classification: Each instance is assigned a label: $(x, l \in L)$
- Multi-label Classification: Each instance is assigned a subset of labels: $(x, S \subseteq L)$
Introduction

- A set of instances: \( D = \{x_0, x_1, \cdots, x_m\} \)
- A set of predefined labels: \( L = \{l_0, l_1, \cdots, l_n\} \)
- Single-label Classification: Each instance is assigned a label: \((x, l \in L)\)
- Multi-label Classification: Each instance is assigned a subset of labels: \((x, S \subseteq L)\)

Example Applications
- a film can be labeled Romance and Comedy
- a news article can be about Science and Technology
- an image can contain Beach, Sunset and Mountains
- a patient’s symptoms may correspond to various ailments
- a collection of genes can have multiple functions
Introduction

- A set of instances: $D = \{x_0, x_1, \cdots, x_m\}$
- A set of *predefined* labels: $L = \{l_0, l_1, \cdots, l_n\}$
- Single-label Classification: Each instance is assigned a label: $(x, l \in L)$
- Multi-label Classification: Each instance is assigned a *subset* of labels: $(x, S \subseteq L)$

**Example Applications**
- a film can be labeled Romance and Comedy
- a news article can be about Science and Technology
- an image can contain Beach, Sunset and Mountains
- a patient’s symptoms may correspond to *various ailments*
- a collection of genes can have *multiple functions*

**Some Multi-label-centric Issues**
- label correlations
  - consider $\{\text{Romance, Comedy}\}$ vs $\{\text{Romance, Horror}\}$
- computational complexity
Problem Transformation

Any multi-label problem can be transformed into one or several single-label problems. Any single-label classifier can be used.

- Problem transformation is core to most multi-label classification, even “algorithm adaption” methods
- There are several “base” methods common to many works
  - e.g. Combination Method (CM)
Problem Transformation

Any multi-label problem can be transformed into one or several single-label problems. Any single-label classifier can be used.

- Problem transformation is core to most multi-label classification, even “algorithm adaption” methods
- There are several “base” methods common to many works
  - e.g. Combination Method (CM)

Combination Method (CM)

Each label subset $S \subseteq L$ is treated as a single label, thus forming a single-label problem. The distinct label sets are the possible single labels.

- takes into account label correlations
- many single labels to choose from
- cannot predict new combinations
The Pruned Sets Method (PS)

- Multi-label data:
  - Some label correlations are very frequent
  - Most label correlations are very infrequent
The Pruned Sets Method (PS)

- Multi-label data:
  - Some label correlations are very frequent
  - Most label correlations are very infrequent

The Pruned Sets Method (PS)

- Treat each label set as a single-label (as per CM)
  - preserves label correlation information
- Prune away infrequent sets and;
- decompose these sets into frequent sets
  - e.g. \((\text{movie}_i, \{\text{Romance}, \text{Comedy}, \text{Horror}\})\) (infrequent)
    \(\rightarrow (\text{movie}_i, \{\text{Romance}, \text{Comedy}\}), (\text{movie}_i, \{\text{Comedy}, \text{Horror}\})\) ...
  - represents only the core label sets as single-labels
  - fewer single labels to learn/choose from (efficient/less error prone)
Multi-label data:
- Some label correlations are very frequent
- Most label correlations are very infrequent

The Pruned Sets Method (PS)

Treat each label set as a single-label (as per CM)
- preserves label correlation information

Prune away infrequent sets and;

decompose these sets into frequent sets
- e.g. \((movie_i, \{\text{Romance}, \text{Comedy}, \text{Horror}\})\) (infrequent)
  \rightarrow (movie_i, \{\text{Romance}, \text{Comedy}\}), (movie_i, \{\text{Comedy}, \text{Horror}\})\ldots
- represents only the core label sets as single-labels
- fewer single labels to learn/choose from (efficient/less error prone)
- cannot predict new combinations
- prone to over-fitting the data
Ensembles of Pruned Sets (EPS)

- Several PS classifiers trained on *subsets* of the training data
  - introduces variation
- The predictions are combined to form *new combinations*
  - reduces over-fitting
  - more robust
Ensembles of Pruned Sets (EPS)

- Several PS classifiers trained on subsets of the training data
  - introduces variation
- The predictions are combined to form new combinations
  - reduces over-fitting
  - more robust

Example (EPS - Classification Phase)

<table>
<thead>
<tr>
<th>Ensemble SL Predictions</th>
<th>PS₀</th>
<th>PS₁</th>
<th>PS₂</th>
<th>PS₃</th>
<th>PS₄</th>
<th>PS₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>(M)</td>
<td>(A,F)</td>
<td>(A,C)</td>
<td>(A,F)</td>
<td>(M)</td>
<td>(M)</td>
<td></td>
</tr>
</tbody>
</table>
Ensembles of Pruned Sets (EPS)

- Several PS classifiers trained on *subsets* of the training data
  - introduces variation
- The predictions are combined to form new combinations
  - reduces over-fitting
  - more robust

Example (EPS - Classification Phase)

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>PS₀</th>
<th>PS₁</th>
<th>PS₂</th>
<th>PS₃</th>
<th>PS₄</th>
<th>PS₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>SL Predictions</td>
<td>(M)</td>
<td>(A,F)</td>
<td>(A,C)</td>
<td>(A,F)</td>
<td>(M)</td>
<td>(M)</td>
</tr>
<tr>
<td>Counts</td>
<td>A</td>
<td>M</td>
<td>F</td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ensembles of Pruned Sets (EPS)

- Several PS classifiers trained on *subsets* of the training data
  - introduces variation
- The predictions are combined to form *new combinations*
  - reduces over-fitting
  - more robust

Example (EPS - Classification Phase)

<table>
<thead>
<tr>
<th>Ensemble</th>
<th>SL Predictions</th>
<th>PS&lt;sub&gt;0&lt;/sub&gt;</th>
<th>PS&lt;sub&gt;1&lt;/sub&gt;</th>
<th>PS&lt;sub&gt;2&lt;/sub&gt;</th>
<th>PS&lt;sub&gt;3&lt;/sub&gt;</th>
<th>PS&lt;sub&gt;4&lt;/sub&gt;</th>
<th>PS&lt;sub&gt;5&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classif. (⊆ L)</td>
<td>(M)</td>
<td>(A, F)</td>
<td>(A, C)</td>
<td>(A, F)</td>
<td>(M)</td>
<td>(M)</td>
<td></td>
</tr>
<tr>
<td>Counts</td>
<td>A</td>
<td>0.375</td>
<td>M</td>
<td>0.375</td>
<td>F</td>
<td>0.250</td>
<td>C</td>
</tr>
</tbody>
</table>

\[ t = 0.2 \]
**Experiments / Results**

- *Reuters* dataset ($|D|=6000$, $|L|=103$) 50/50 train/test split
- **BM**: Binary Method (one binary classifier per label)
- **CM**: Combination Method (each set is a single-label)
- **EPS, RAKEL**: 10 models, auto-tuned threshold, varying $p,k$
  - e.g. $p=3$: only label sets occurring $>3$ times are *frequent*
- *All using Support Vector Machines as single-label classifiers*

<table>
<thead>
<tr>
<th></th>
<th>BM</th>
<th>EPS</th>
<th>RAKEL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>Acc.</td>
<td>$p$</td>
</tr>
<tr>
<td><strong>BM</strong></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>123</td>
<td>32.48</td>
<td>4</td>
</tr>
<tr>
<td>CM</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1,379</td>
<td>48.75</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>RAKEL</strong></td>
<td>$k$</td>
<td>Time</td>
<td>Acc.</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>10</td>
<td>10.05</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>350</td>
<td>36.66</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>3,627</td>
<td>44.70</td>
</tr>
<tr>
<td>61*</td>
<td></td>
<td>22,337</td>
<td>47.35</td>
</tr>
<tr>
<td>102</td>
<td></td>
<td>DNF</td>
<td>DNF</td>
</tr>
</tbody>
</table>
Conclusions

- Ensembles of Pruned Sets: A new problem transformation method
  - classifier independent
  - improved performance over BM, CM, and RAKEL
  - efficient in practice
- Main contribution: focus on core label correlations
  - pruning infrequent sets
  - set decomposition into frequent sets
  - flexible pruning parameter $p$
  - can be combined easily with other methods
### Description.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scene</strong></td>
<td>2407</td>
<td>6</td>
<td>1.07</td>
<td>0.006</td>
<td>still scenes</td>
</tr>
<tr>
<td><strong>Yeast</strong></td>
<td>2417</td>
<td>14</td>
<td>4.24</td>
<td>0.082</td>
<td>protein function</td>
</tr>
<tr>
<td><strong>Medical</strong></td>
<td>978</td>
<td>45</td>
<td>1.25</td>
<td>0.096</td>
<td>medical text</td>
</tr>
<tr>
<td><strong>Enron</strong></td>
<td>1702</td>
<td>53</td>
<td>3.38</td>
<td>0.442</td>
<td>e-mail corpus</td>
</tr>
<tr>
<td><strong>Reuters</strong></td>
<td>6000</td>
<td>103</td>
<td>1.46</td>
<td>0.147</td>
<td>newswire stories</td>
</tr>
</tbody>
</table>

- $D = $ full dataset
- $L = $ label set
- $LC = $ Label Cardinality. Average number of labels per instance in $D$
- $PD = $ Percent $D$instinct. The percentage of instances with a distinct label set
Framework

- WEKA\(^1\) framework
- using Support Vector Machines (SVM) as single-label classifiers (default parameters)
- 5 × 2 Cross Validation (CV)

Problem Transformation parameters

- trialled in order according to theoretical complexity
- under 5 × CV on training set
- cut off: 1 hour per parameter combination

Evaluation Methods

- Accuracy\((D)\) = \(\frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|S_i \cap Y_i|}{|S_i \cup Y_i|}\)
- Micro \(F_1(D)\) = \(\frac{1}{|D|} \sum_{i=1}^{|D|} \frac{2 \times \text{prec}_i \times \text{recall}_i}{\text{prec}_i + \text{recall}_i}\)
- Hamming loss\((D)\) = \(1 - \frac{1}{|D| \times |L|} \sum_{i=1}^{|D|} |S_i \oplus Y_i|\)

\(^1\)http://www.cs.waikato.ac.nz/ml/weka/
CM: Combination Method
BM: Binary Method
RM: Ranking Method

- Tune threshold $t = \{0.1, \cdots, 0.9\}$

PS: Pruned Sets method

- Tune parameter $p = \{5, 4, 3, 2, 1\}$
- Tune parameter $s = \{-, A_1, A_2, A_3, B_1, B_2, B_3\}$

EPS: Ensembles of Pruned Sets

- Tune parameters using a single PS method
- Tune threshold $t = \{0.1, \cdots, 0.9\}$

RAKEL: RAndom K labEL subsets

- Parameter range as per paper
- Tune threshold $t = \{0.1, \cdots, 0.9\}$
Accuracy Measure

Paired $t$ Test (against CM)

$\uparrow, \downarrow$ statistically significant improvement, degradation
<table>
<thead>
<tr>
<th></th>
<th>BM</th>
<th>[CM]</th>
<th>RAKEL</th>
<th>PS</th>
<th>EPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene</td>
<td>0.671 \downarrow</td>
<td>0.729</td>
<td>0.735</td>
<td>0.730</td>
<td>0.752 \uparrow</td>
</tr>
<tr>
<td>Yeast</td>
<td>0.630</td>
<td>0.633</td>
<td>0.664 \uparrow</td>
<td>0.643</td>
<td>0.655 \uparrow</td>
</tr>
<tr>
<td>Medical</td>
<td>0.791 \uparrow</td>
<td>0.767</td>
<td>0.784</td>
<td>0.766</td>
<td>0.764</td>
</tr>
<tr>
<td>Enron</td>
<td>0.504</td>
<td>0.502</td>
<td>0.543 \uparrow</td>
<td>0.520</td>
<td>0.543 \uparrow</td>
</tr>
<tr>
<td>Reuters</td>
<td>0.421 \downarrow</td>
<td>0.482</td>
<td>0.418 \downarrow</td>
<td>0.496</td>
<td>0.499 \uparrow</td>
</tr>
</tbody>
</table>

- **$F_1$ Measure**
- **Paired $t$ Test** (against [CM])
  - $\uparrow, \downarrow$ statistically significant improvement, degradation