Outline

• Task Settings
• Previous Work
• Proposed Method
• Results and Discussion
• Conclusion
Outline

• Task Settings
• Previous Work
• Proposed Method
• Results and Discussion
• Conclusion
An increase in the amount of harvested wood products (HWP) from sustainable forestry would help to reduce levels of atmospheric carbon. In the first commitment period of the Kyoto Protocol (2008-2012), this carbon stock effect of HWP is ignored, and forest harvesting is treated as an instantaneous emission of carbon dioxide. However, in the next commitment period of the United Nations Framework Convention on Climate Change from 2013, the carbon stock changes resulting from HWP will be taken into account in the national greenhouse gas inventories...
**Task: Multi-Label Text Classification**

<table>
<thead>
<tr>
<th>Input: Document</th>
<th>Output: A set of labels</th>
</tr>
</thead>
</table>
| **Journal of Wood Science** | **• FF01020X (forestry policies)**  
| Harvested wood products accounting in the post Kyoto commitment period | **• SA01020V (environmental problems)** |
| An increase in the amount of harvested wood products (HWP) from sustainable forestry would help to reduce levels of atmospheric carbon. In the first commitment period of the Kyoto Protocol (2008-2012), this carbon stock effect of HWP is ignored, and forest harvesting is treated as an instantaneous emission of carbon dioxide. However, in the next commitment period of the United Nations Framework Convention on Climate Change from 2013, the carbon stock changes resulting from HWP will be taken into account in the national greenhouse gas inventories... |
Task: Hierarchically Organized Labels

- FF01010M: forestry - general
- FF01020X: forestry policies
- FF01030I: environmental problems
- SA01020V: environmental problems
- SA01030G: environmental problems
Outline

• Task Settings
• Previous Work
• Proposed Method
• Results and Discussion
• Conclusion
Question: What is the Hierarchy for Text Classification?
Question: What is the Hierarchy for Text Classification?

1. Flat: Nothing. Just ignore the hierarchy.
Question: What is the Hierarchy for Text Classification?

1. Flat: Nothing. Just ignore the hierarchy.
2. Pruning: Top-down pruning for speed, often in exchange for accuracy
Question: What is the Hierarchy for Text Classification?

1. Flat: Nothing. Just ignore the hierarchy.
2. Pruning: Top-down pruning for speed, often in exchange for accuracy.
3. Parameter Sharing: Share parameters according to the hierarchy to improve accuracy.
Question: What is the Hierarchy for Text Classification?

1. Flat: Nothing. Just ignore the hierarchy.
2. Pruning: Top-down pruning for speed, often in exchange for accuracy
3. Parameter Sharing: Share parameters according to the hierarchy to improve accuracy
4. **Proposed Method**: To capture dependencies among multiple labels to be output
Building Block: Binary Linear Classifier 1/2

- Document
  - Harvested wood products accounting...
- Binary classifier for forestry
- Parameters
  - +1 (Yes)
  - -1 (No)
Building Block: Binary Linear Classifier 1/2

Harvested wood products accounting...

(training data, -1)
(training data, -1)
(training data, +1)

training

document

parameters

binary classifier for forestry

+1 (Yes) or -1 (No)
Building Block: Binary Linear Classifier 1/2

document

Harvested wood products accounting...
Building Block: Binary Linear Classifier 1/2

document

Harvested wood products accounting...

feature vector

\[
\begin{pmatrix}
0 \\
0 \\
1 \\
1 \\
0 \\
0 \\
0 \\
\vdots
\end{pmatrix}
\]
Building Block: Binary Linear Classifier 1/2

Document:
Harvested wood products accounting...

Feature vector:
\[
\begin{bmatrix}
0 \\
0 \\
1 \\
1 \\
0 \\
0 \\
0 \\
0 \\
1 \\
1 \\
0 \\vdots
\end{bmatrix}
\]
- Algorithm
- Dolphin
- Wood
- Products
Building Block: Binary Linear Classifier 1/2

Document:
Harvested wood products accounting...

Feature vector:

weight vector:
(-0.5, -0.2, 0.8, ...)

(weight vector)
Building Block: Binary Linear Classifier 1/2

Document:

Harvested wood products accounting...

Feature vector:

\[
\begin{bmatrix}
0 \\
0 \\
1 \\
1 \\
0 \\
0 \\
0 \\
1 \\
\vdots
\end{bmatrix}
\]

- algorithm
- dolphin
- wood products

Weight vector:

\((-0.5, -0.2, 0.8, \ldots)\)

If \(>0\) then +1, otherwise -1
Building Block: Binary Linear Classifier 1/2

- Harvested wood products accounting...
- Document
- Weight vector: $(-0.5, -0.2, 0.8, \ldots)$
- Feature vector:
  
  \[
  \begin{bmatrix}
  0 \\
  0 \\
  1 \\
  1 \\
  0 \\
  0 \\
  0 \\
  0 \\
  1 \\
  0 \\
  \vdots
  \end{bmatrix}
  \]

- Algorithm dolphin wood products
- $>0$ ? then +1
- otherwise -1

This is the parameters to be learned
Flat Model (ignores label hierarchy)

set of binary classifiers
Flat Model (ignores label hierarchy)

parameters for FF0102X

FF01020X ?
FF01030I ?
FF01040T ?
\vdots

set of binary classifiers
Flat Model (ignores label hierarchy)

original training data

(set, FF01020X, SA01020V)
(set, FF0510O)
(set, FF01020X)
(set, SA01020V, QJ05023V)

set of binary classifiers

parameters for FF0102X

FF01020X ?
FF01030I ?
FF01040T ?
...
Flat Model (ignores label hierarchy)

training data for FF01020X

original training data

parameters for FF0102X

set of binary classifiers
Tree Models 1/2

Binary classifier for each edge
Tree Models 2/2

original training data

(parameters for F→FF01)

(⋯, FF01020X, SA01020V)
(⋯, FF05100)
(⋯, FF01020X)
(⋯, SA01020V, QJ05023V)
⋯
Tree Models 2/2

training data for F→FF01:

Original training data:

- ( , +1)
- ( , -1)
- ( , +1)
- ( , -1)

...
This document is irrelevant to $F$. Use this document for training?

- Yes → All data method
- No → Sibling data method
Top-down Pruning
Top-down Pruning
Top-down Pruning
Top-down Pruning
Top-down Pruning
Top-down Pruning
Outline

• Task Settings
• Previous Work
• Proposed Method
• Results and Discussion
• Conclusion
Proposed Method: Hierarchy for Multiple Labels

• Need to choose not just one best
• How many labels should we choose? A difficult question even for humans
• Our assumption: Human annotators do not select each label independently but consider the relative importance among competing labels by consulting the hierarchy
Intuition: Hierarchy Encodes Inter-label Dependencies
Intuition: Hierarchy Encodes Inter-label Dependencies

We need both; Both are relevant.
Intuition: Hierarchy Encodes Inter-label Dependencies

Both are reasonably relevant but choosing both is bad. They are **redundant**.

We need both; Both are relevant.
Intuition: Hierarchy Encodes Inter-label Dependencies

Both are reasonably relevant but choosing both is bad. They are **redundant**.

We need both; Both are relevant.

The competitive nature is reflected in the hierarchy.
Preparation: Global Model

- Want to capture dependencies among multiple labels to be output
- First need to jointly predicts multiple labels
- Then capture inter-label dependencies by adding features to the global model
Global Model: Find a Subtree that maximizes the sum of scores
Global Model: Find a Subtree that maximizes the sum of scores
Global Model: Find a Subtree that maximizes the sum of scores
Global Model: Find a Subtree that maximizes the sum of scores
Global Model: Find a Subtree that maximizes the sum of scores

![Tree Diagram]

- **A**
- **F**
- **S**
- **Z**

**Root Node:**
- **FF01**
- **FF02**
- **FF03**

**Leaf Nodes:**
- **SA01**
- **SA02**
- **SA03**
- **SA04**

Scores:
- **1.5**
- **-0.9**
- **1.2**
- **-1.6**
- **1.5**
- **-0.2**
- **-0.9**
- **-0.2**
Global Model: Find a Subtree that maximizes the sum of scores
Global Model: Find a Subtree that maximizes the sum of scores
Global Model: Find a Subtree that maximizes the sum of scores
Global Model: Find a Subtree that maximizes the sum of scores

-2.6 \rightarrow 0.9 \rightarrow 4.7 \rightarrow 2.7 \rightarrow 0.1

0.5 \rightarrow -0.2 \rightarrow 0.2 \rightarrow 1.5

0 \rightarrow 0 \rightarrow 0 \rightarrow 0

0 \rightarrow 0 \rightarrow 0 \rightarrow 0
Global Model: Find a Subtree that maximizes the sum of scores

Dynamic programming finds an exact solution
Global Model: Find a Subtree that maximizes the sum of scores

Dynamic programming finds an exact solution

Much slower than top-down pruning

Much slower than top-down pruning
Additonal Scores by Branching Features

![Diagram of branching features with scores]

- FF01: 0.03
- FF02: 0.2
- FF03: -0.1
- SA01: -0.4
Additonal Scores by Branching Features

Select $N$ children:
- $N=1$ -0.05
- $N=2$ -0.10
- $N=3$ -0.20
- $N>3$ -0.25
Additonal Scores by Branching Features

Select $N$ children:
- $N=1$ -0.05
- $N=2$ -0.10
- $N=3$ -0.20
- $N>3$ -0.25
Additonal Scores by Branching Features

Select $N$ children:
- $N=1$ -0.05
- $N=2$ -0.10
- $N=3$ -0.20
- $N>3$ -0.25
Select $N$ children:
- $N=1$ : -0.02
- $N=2$ : -0.05
- $N=3$ : -0.10
- $N>3$ : -0.15

Select $N$ children:
- $N=1$ : -0.05
- $N=2$ : -0.10
- $N=3$ : -0.20
- $N>3$ : -0.25

The diagram shows a branching structure with scores assigned to each branch. The scores range from -2.6 to 1.2.
Select $N$ children:
- $N=1$  -0.02
- $N=2$  -0.05
- $N=3$  -0.10
- $N>3$  -0.15

Select $N$ children:
- $N=1$  -0.05
- $N=2$  -0.10
- $N=3$  -0.20
- $N>3$  -0.25

Dynamic programming still applicable
Training

• Global optimization
  – Instead of training edge classifiers separately, we directly optimize the global model

• Parallelization by iterative parameter mixing (McDonald+, 2010)
  – Global optimization is too slow
  – Parallelize training by splitting training data into small “shards” and mix the models per iteration
Outline

• Task Settings
• Previous Work
• Proposed Method
• Results and Discussion
• Conclusion
Experimental Settings

• Portion of JSTPlus (Japanese academic bibliographic database)
  – Training: 409,892 documents
  – Evaluation: 45,419 documents

• Features: content words & journal name

• Training:
  – Online passive-aggressive algorithm
  – 10 iterations
Example-based F Measure

- Flat: 0.40
- Top-down (all data): 0.37
- Top-down (sib data): 0.38
- Global: 0.43
- Global + Branching: 0.43

EBF
Label-based Micro-average F Measure

- Flat: 0.39
- Top-down (all data): 0.37
- Top-down (sib data): 0.37
- Global: 0.43
- Global + Branching: 0.43
Model Size

<table>
<thead>
<tr>
<th>Method</th>
<th># of elems in the weight vector (Million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat</td>
<td>73</td>
</tr>
<tr>
<td>Top-down (all data)</td>
<td>115</td>
</tr>
<tr>
<td>Top-down (sib data)</td>
<td>39</td>
</tr>
<tr>
<td>Global</td>
<td>68</td>
</tr>
<tr>
<td>Global + Branching</td>
<td>62</td>
</tr>
</tbody>
</table>
Branching features reduce the model size while slightly improving accuracy.
Conclusions & Future Work

• Exploit the label hierarchy to capture inter-label dependencies
• Branching features reduce model size while slightly improving accuracy
• Future work
  – Extension to directed acyclic graphs
  – Improving scalability