Learning Context-dependent Label Permutations for Multi-label Classification

Jinseok Nam
Amazon Alexa AI

Joint work with
Young-Bum Kim, Eneldo Loza Mencía, Sunghyun Park, Ruhi Sarikaya and Johannes Fürnkranz
Multi-label Classification (MLC)

- **Goal**: learn a function $f$ that maps instances to a subset of labels

It is important to take into account **label dependencies**.

Joint probability of labels

$$P(y_1, y_2, \cdots, y_L | x) = \prod_{i=1}^{L} P(y_i | y_{<i}, x)$$
Maximization of the joint probability

• Traditional approaches for minimizing subset 0/1 loss:
  • (Probabilistic) classifier chain (Dembczyński et al., ICML 2010; Read et al., MLJ 2011)

\[ Y = \{ \text{Sea, Desert, Building, Sky, Cloud, Mountain} \} \]

1. Creates a chain of \( L \) labels
2. Train \( L \) independent classifiers given input and partial label vector

Additional input features
Maximization of the joint probability

- Traditional approaches for minimizing **subset 0/1 loss:**
  - (Probabilistic) classifier chain

\[ Y = \{ \text{Sea, Desert, Building, Sky, Cloud, Mountain} \} \]

1. Creates a chain of \( L \) labels

2. Train \( L \) independent classifiers given input and partial label vector

**Additional input features**

**Limitations**
- Error-propagation at test time
- Effect of label orders in the chain
Maximization of the joint probability

- Traditional approaches for minimizing **subset 0/1 loss**: (Dembczyński et al., ICML 2010; Read et al., MLJ 2011)

\[ Y = \{\text{Sea, Desert, Building, Sky, Cloud, Mountain}\} \]

1. Creates a chain of \( L \) labels

\[
\begin{align*}
\text{Desert} & \quad \text{Sea} & \quad \text{Cloud} & \quad \text{Mountain} & \quad \text{Sky} & \quad \text{Building} \\
\quad f_1 & \quad f_2 & \quad f_3 & \quad f_4 & \quad f_5 & \quad f_6 \\
\end{align*}
\]

- Additional input features
- 2. Train \( L \) independent classifiers given input and partial label vector

**Limitations**

- Error-propagation at test time
- Effect of label orders in the chain
Recurrent Neural Networks for MLC

- Learning from a set of relevant labels in a sequential manner (Nam et al., NIPS 2017)
  - Number of relevant labels is much smaller than the total number of labels
Recurrent Neural Networks for MLC

- Learning from a set of relevant labels in a sequential manner (Nam et al., NIPS 2017)
  - Number of relevant labels is much smaller than the total number of labels

• Question: The effect of label permutation remain!
  How to determine the target label permutation?
Target label permutations for RNN training

• Static label permutation for all instances
  • Arbitrary label sequence randomly chosen at the beginning
  • Label frequency distribution: $freq2rare$, $rare2freq$
  • Label structures (e.g., pairwise label dependencies)

→ Suboptimal choice; learn from only one permutation
Target label permutations for RNN training

- Static label permutation for **all** instances
  - Arbitrary label sequence randomly chosen at the beginning
  - Label frequency distribution: $freq2rare$, $rare2freq$
  - Label structures (e.g., pairwise label dependencies)

  → *Suboptimal* choice; learn from only one permutation

- Different label permutations for **individual** instances
  - Choosing randomly every time
  - Learning from all possible label permutations

  → More robust to the effect of label permutation; *Computational complexity*
Target label permutations for RNN training

• Static label permutation for *all* instances
  • Arbitrary label sequence randomly chosen at the beginning
  • Label frequency distribution: freq2rare, rare2freq
  • Label structures (e.g., pairwise label dependencies)

→ *Suboptimal* choice; learn from only one permutation

• Different label permutations for *individual* instances
  • Choosing randomly every time
  • Learning from all possible label permutations

→ More robust to the effect of label permutation; *Computational complexity*

We need MLC algorithms that learn context-dependent label permutations *efficiently*!
Model based label permutation

(1) label sequence sampling

(2) computing errors & updating parameters

true target label set: 1 2 3 4 5

sampled target label permutation: 2 1 4 3 5

False positive prediction
True positive prediction
False negative prediction
\[ \nabla_{\theta} J(\theta) = \mathbb{E}_{P_{\theta}^T} \left[ \sum_{i=0}^{T-1} \nabla_{\theta} \log P_{\theta}(a_i|s_i)(R_i - b(s_i)) \right] \]
Experiments

- We combined two approaches! Context-dependent label permutation learning clearly outperforms static label permutation approaches

<table>
<thead>
<tr>
<th>Methods</th>
<th>Example $F_1$</th>
<th>Macro $F_1$</th>
<th>Prec@1</th>
<th>Prec@3</th>
<th>Prec@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLEEC</td>
<td>-</td>
<td>-</td>
<td>87.82</td>
<td>73.45</td>
<td>59.17</td>
</tr>
<tr>
<td>FastXML</td>
<td>-</td>
<td>-</td>
<td>84.22</td>
<td>67.33</td>
<td>53.04</td>
</tr>
<tr>
<td>Parabel</td>
<td>-</td>
<td>-</td>
<td>83.91</td>
<td>67.12</td>
<td>52.99</td>
</tr>
<tr>
<td>freq2rare</td>
<td>66.63±0.33</td>
<td>39.68±0.69</td>
<td>90.05±0.31</td>
<td>74.20±0.18</td>
<td>58.39±0.29</td>
</tr>
<tr>
<td>rare2freq</td>
<td>66.95±0.26</td>
<td>43.33±0.62</td>
<td>53.67±1.31</td>
<td>59.57±0.78</td>
<td>52.49±0.47</td>
</tr>
<tr>
<td>fixed-rd</td>
<td>67.21±0.25</td>
<td>41.85±0.90</td>
<td>73.95±5.20</td>
<td>65.58±2.31</td>
<td>55.55±0.83</td>
</tr>
<tr>
<td>always-rd</td>
<td>66.25±0.25</td>
<td>34.03±0.58</td>
<td>89.08±0.18</td>
<td>73.90±0.24</td>
<td>59.45±0.31</td>
</tr>
<tr>
<td>CLP-RNN ($\alpha=0$)</td>
<td>67.22±0.15</td>
<td>38.75±0.88</td>
<td>89.40±0.42</td>
<td>73.84±0.30</td>
<td>59.29±0.17</td>
</tr>
<tr>
<td>CLP-RNN ($\alpha=0.6$)</td>
<td>67.27±0.30</td>
<td>36.49±0.74</td>
<td>91.27±0.28</td>
<td>75.25±0.32</td>
<td>59.75±0.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Example $F_1$</th>
<th>Macro $F_1$</th>
<th>Prec@1</th>
<th>Prec@3</th>
<th>Prec@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLEEC</td>
<td>-</td>
<td>-</td>
<td>67.59</td>
<td>61.38</td>
<td>56.56</td>
</tr>
<tr>
<td>FastXML</td>
<td>-</td>
<td>-</td>
<td>69.61</td>
<td>64.12</td>
<td>59.27</td>
</tr>
<tr>
<td>Parabel</td>
<td>-</td>
<td>-</td>
<td>67.44</td>
<td>61.83</td>
<td>56.75</td>
</tr>
<tr>
<td>freq2rare</td>
<td>31.36±0.17</td>
<td>13.94±0.29</td>
<td>57.21±0.38</td>
<td>54.28±0.31</td>
<td>51.16±0.36</td>
</tr>
<tr>
<td>rare2freq</td>
<td>31.60±0.15</td>
<td>18.00±0.31</td>
<td>17.46±0.38</td>
<td>18.49±0.51</td>
<td>20.31±0.72</td>
</tr>
<tr>
<td>fixed-rd</td>
<td>32.74±0.27</td>
<td>16.48±0.31</td>
<td>40.59±1.31</td>
<td>37.21±3.06</td>
<td>35.74±2.60</td>
</tr>
<tr>
<td>always-rd</td>
<td>32.45±0.05</td>
<td>13.00±0.23</td>
<td>66.58±0.99</td>
<td>60.46±0.54</td>
<td>54.95±0.55</td>
</tr>
<tr>
<td>CLP-RNN ($\alpha=0$)</td>
<td>34.43±0.54</td>
<td>17.33±0.17</td>
<td>69.57±1.43</td>
<td>61.57±0.69</td>
<td>55.73±0.56</td>
</tr>
<tr>
<td>CLP-RNN ($\alpha=0.9$)</td>
<td>35.80±0.35</td>
<td>18.00±0.51</td>
<td>70.54±0.77</td>
<td>63.39±0.65</td>
<td>57.72±0.58</td>
</tr>
</tbody>
</table>