Implicit Learning of Simpler Output Kernels for Multi-label Prediction

Hanchen Xiong, Sandor Szedmak, Justus Piater
UNIVERSITY OF INNSBRUCK, AUSTRIA

Overview

Research in Multi-label Prediction:

- exploiting and utilizing inter-label dependencies;
- increasingly more sophisticated dependency is used?
- “overfit” output structural dependencies when the desired ones are simpler
- a regularization should be added on output structural dependencies.

Our contributions: Joint SVM

- joint SVM ↔ structural SVM with linearly decomposable score functions;
- in joint SVM, (non-)linear kernels can be assigned on outputs to capture inter-label dependencies;
- joint SVM shares the same computational complexity as a single SVM;
- when linear output kernels are used, a output-kernel regularization is implicitly added.
- yield promising results on 3 image annotation databases.

From Structural SVM to Joint SVM

Structural SVM is an extension of SVM for structured-outputs, in which, however, the margin to be maximized is defined as the score gap between the desired output and the first runner-up.

\[
\arg \min_{W \in \mathbb{R}^{n \times m}} \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{m} \max_{y \in \mathbb{I}} \left\{ d(y^{(i)}, y^*) - \Delta_F(y^{(i)}, y^*) \right\}
\]

where \(\Delta_F(y^{(i)}, y^*) = F(x^{(i)}, y^{*}; W) - F(x^{(i)}, y^{*}; W)\), the score function is \(F(x^{(i)}, y; W) = \langle W, \phi(x^{(i)}) \rangle \otimes y^{*}\), and Hamming distance \(d(y^{(i)}, y^*)\) is used on outputs. Of linear decomposability, (I) can be rewritten as:

\[
\begin{align*}
\arg \min_{W \in \mathbb{R}^{n \times m}} & \quad \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{m} \max_{y \in \mathbb{I}} \left\{ d(y^{(i)}, y^*) - \Delta_F(y^{(i)}, y^*) \right\} \\
\text{s.t.} & \quad \langle W, \phi(x^{(i)}) \rangle \otimes y^{*} \geq 1 - \xi^{(i)}, \xi^{(i)} \geq 0, i \in \{1, \ldots, m\}
\end{align*}
\]

where \(\langle \cdot, \cdot \rangle\) denotes Frobenius product and \(\|W\|\) is the Frobenius norm of matrix \(W\).

It can be seen in (2) that, with linearly decomposable score functions and output dependencies, structural SVM on multi-label learning is equivalent to learning \(T\) SVMs jointly: Joint SVM

\[
\begin{align*}
\arg \min_{\theta \in \mathbb{R}^{n \times m}} & \quad \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{m} \xi^{(i)} \\
\text{s.t.} & \quad \langle \phi\left(y^{(i)}\right), \phi(x^{(i)}) \rangle \geq 1 - \xi^{(i)}, \xi^{(i)} \geq 0, i \in \{1, \ldots, m\}
\end{align*}
\]

where \(y^{(i)} = [y^{(i)}_1, \ldots, y^{(i)}_m], W = \left[\frac{w_1^T}{\sqrt{m}}, \ldots, \frac{w_m^T}{\sqrt{m}}\right]^T\). The corresponding dual form of Joint SVM is:

\[
\begin{align*}
\arg \min_{\alpha_1, \ldots, \alpha_m} & \quad \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j K_{\phi}(y^{(i)}, y^{(j)})K_{\phi}(x^{(i)}, x^{(j)}) \\
\text{s.t.} & \quad \forall i, 0 \leq \alpha_i \leq C
\end{align*}
\]

therefore, the same computational complexity as a single regular SVM.

Implicit Regularization on Linear Output Kernels

When a linear output kernel is used to capture pairwise dependencies of labels, via which the output vectors can be linearly mapped as \(\psi(y) = Py: K_{\phi_{lin}}(y^{(i)}, y^{(j)}) = y^{(i)\top} \Omega y^{(j)}\) where \(\Omega = P^\top P\). By denoting \(U = P^\top W\), we can have:

\[
\begin{align*}
\arg \min_{W \in \mathbb{R}^{n \times m}} & \quad \frac{1}{2} \|W\|^2 + C \sum_{i=1}^{m} \xi^{(i)} \\
\text{s.t.} & \quad \langle \psi(y^{(i)}), \psi(x^{(i)}) \rangle \geq 1 - \xi^{(i)}, \xi^{(i)} \geq 0, i \in \{1, \ldots, m\}
\end{align*}
\]

we use a compact regularization for both \(W\) and \(\Omega\): \(\frac{1}{2} \|W\|^2 + C \sum_{i=1}^{m} \xi^{(i)}\) resulting in:

\[
\begin{align*}
\arg \min_{U \in \mathbb{R}^{n \times m}} & \quad \frac{1}{2} \|U\|^2 + C \sum_{i=1}^{m} \xi^{(i)} \\
\text{s.t.} & \quad \langle \psi(U), \psi(U) \rangle \geq 1 - \xi^{(i)}, \xi^{(i)} \geq 0, i \in \{1, \ldots, m\}
\end{align*}
\]

Remarkably, a linear output kernel is implicitly learned, and absorbed in \(W\), also a regularization on the output kernel is also implicitly added.

Experimental Results and Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Corel5K</th>
<th>ESPgame</th>
<th>IAPRTC12</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P (%)</td>
<td>R (%)</td>
<td>F (%)</td>
</tr>
<tr>
<td>MBMRM</td>
<td>24.0</td>
<td>25.0</td>
<td>24.0</td>
</tr>
<tr>
<td>JEC</td>
<td>27.0</td>
<td>32.0</td>
<td>29.0</td>
</tr>
<tr>
<td>TagProp</td>
<td>33.0</td>
<td>42.0</td>
<td>37.0</td>
</tr>
<tr>
<td>FacTag</td>
<td>43.0</td>
<td>43.0</td>
<td>43.0</td>
</tr>
<tr>
<td>SVM</td>
<td>48.5</td>
<td>38.0</td>
<td>42.6</td>
</tr>
<tr>
<td>SVM+Pol(2)</td>
<td>46.6</td>
<td>37.0</td>
<td>41.3</td>
</tr>
<tr>
<td>SVM+Pol(3)</td>
<td>41.5</td>
<td>31.3</td>
<td>35.7</td>
</tr>
</tbody>
</table>

Table 1: Comparison between different versions of joint SVM and other related methods on three benchmark datasets. P, R and F1 denote precision, recall and F1 measure respectively.