A Hassle-Free Unsupervised Domain Adaptation Method Using Instance Similarity Features

Written by Jianfei Yu and Jing Jiang for ACL 2015
Presented by Xia Cui for NLP@UoL
Main Idea

• Using Instance Similarity Features
  • Use unlabelled target domain instances to induce a set of instance similarity features
  • Combine with original features to represent labelled source domain instances
Adaptation with Similarity Features

• Randomly select a subset of target instances from target domain and normalize them (exemplar vector $\mathcal{E} = \{e^{(k)}\}_{k=1}^K$)
• Transform each source instance $x$ into a new feature vector by computing its similarity with each $e^{(k)}$

$$g(x) = [s(x, e^{(1)}), \ldots, s(x, e^{(K)})]^T$$

  • $s(x, x')$: similarity function between $x$ and $x'$ (cosine similarity)
  • Transformed into a K-dimensional vector
• Append this vector to the original feature vector of source instance and use the combined feature vectors of all labelled source instances to train a classifier
Learning in the Target Subspace

• The “hope” of UDA is to “couple” the learning of weights for target-specific features with that of common features
• Input vector space: typically high-dimensional for NLP tasks
• Can be lower dimension because of the strong feature dependence
Learning in the Target Subspace (Cont.)

- Project any input vector \( x \) into three subspaces:
  \[
  x = x_{s,t} + x_{s,\perp} + x_{\perp,t}
  \]
  - \( x_{s,t} \): \( X_s \cap X_T \)
  - \( x_{s,\perp} \): the subspace that is orthogonal to \( x_{s,t} \) spans \( X_s \) that is \( X_{s,\perp} + x_{s,t} = X_s \)
  - \( x_{\perp,t} \): \( X_{\perp,t} + x_{s,t} = X_t \)

- Linear Classifier \( w \) can be decomposed into \( w_{s,t} \), \( w_{s,\perp} \) and \( w_{\perp,t} \):
  \[
  w^\top x = w_{s,t}^\top x_{s,t} + w_{s,\perp}^\top x_{s,\perp} + w_{\perp,t}^\top x_{\perp,t}.
  \]
  \[
  e^{(k)} = e_{s,t}^{(k)} + e_{\perp,t}^{(k)}.
  \]
  \[
  \overline{w} = \sum_k w_k' e_{s,t}^{(k)} + \sum_k w_k' e_{\perp,t}^{(k)}.
  \]
Analysis

• The learned classifier $\bar{w}$ doesn’t have any components in the subspace $X_{s,\perp}$
  • Good! Such a component isn’t useful for T

• The learned $\bar{w}_{\perp,t}$ will unlikely be zero because its learning is “coupled” with the learning of $\bar{w}_{s,t}$ through $w'$

• Pick up target specific features that correlate with useful common features
Reduction of Domain Divergence

• Use a hypothesis space that achieve low error on S while at the same time making it hard to separate S and T instances

<table>
<thead>
<tr>
<th>features</th>
<th>$\hat{\varepsilon}_s$</th>
<th>domain separation error</th>
<th>$\hat{\varepsilon}_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0.000</td>
<td>0.011</td>
<td>0.283</td>
</tr>
<tr>
<td>ISF-</td>
<td>0.120</td>
<td>0.129</td>
<td>0.315</td>
</tr>
<tr>
<td>ISF</td>
<td>0.006</td>
<td>0.062</td>
<td>0.254</td>
</tr>
</tbody>
</table>
Difference from EA++ (Easy Domain Adaption semi-supervised ver.)

• Only learn a single classifier using labelled data from S
  • In EA++, unlabelled target data is used to construct a regularizer that brings the two classifiers of two domains closer
  • In EA++, doesn't restrict either S classifier or T classifier to lie in target subspace
Experiments

• Personalized Spam Filtering (Spam): ECML/PKDD 2006
  • 4000 labelled trained (2500 emails)

• Gene Name Recognition (NER): BioCreAtIve Task 1B
  • Three sets of Medline abstracts with labelled gene names
  • Each set corresponds to a single species
  • DA from one species to another

• Relation Extraction (Relation): ACE 2005
  • Broadcast news and conversational telephone speech
  • 7 major relation types
# Results

<table>
<thead>
<tr>
<th>Method</th>
<th>u00</th>
<th>u01</th>
<th>u02</th>
<th>average</th>
<th>f→y</th>
<th>f→m</th>
<th>m→y</th>
<th>m→f</th>
<th>y→f</th>
<th>y→m</th>
<th>average</th>
<th>Relation average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.678</td>
<td>0.710</td>
<td>0.816</td>
<td>0.735</td>
<td>0.396</td>
<td>0.379</td>
<td>0.526</td>
<td>0.222</td>
<td>0.050</td>
<td>0.339</td>
<td>0.319</td>
<td>0.398</td>
</tr>
<tr>
<td>Common</td>
<td>0.697</td>
<td>0.732</td>
<td>0.781</td>
<td>0.737</td>
<td>0.409</td>
<td>0.388</td>
<td>0.559</td>
<td>0.208</td>
<td>0.059</td>
<td>0.344</td>
<td>0.328</td>
<td>0.401</td>
</tr>
<tr>
<td>SCL</td>
<td>0.699</td>
<td>0.717</td>
<td>0.824</td>
<td>0.747</td>
<td>0.405</td>
<td>0.380</td>
<td>0.525</td>
<td>0.239</td>
<td>0.063</td>
<td>0.35</td>
<td>0.327</td>
<td>0.403</td>
</tr>
<tr>
<td>ISF</td>
<td>0.720</td>
<td>0.769</td>
<td>0.884</td>
<td>0.791**</td>
<td>0.415</td>
<td>0.395</td>
<td>0.566</td>
<td>0.212</td>
<td>0.079</td>
<td>0.360</td>
<td>0.338**</td>
<td>0.416**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>bc→bn</th>
<th>bc→cts</th>
<th>bc→nw</th>
<th>bc→un</th>
<th>bc→wl</th>
<th>bn→bc</th>
<th>bn→cts</th>
<th>bn→nw</th>
<th>bn→un</th>
<th>bn→wl</th>
<th>cts→bc</th>
<th>cts→bn</th>
<th>cts→nw</th>
<th>cts→un</th>
<th>cts→wl</th>
<th>nw2bc</th>
<th>nw→bn</th>
<th>nw→cts</th>
<th>nw→un</th>
<th>nw→wl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.455</td>
<td>0.400</td>
<td>0.445</td>
<td>0.376</td>
<td>0.397</td>
<td>0.528</td>
<td>0.430</td>
<td>0.482</td>
<td>0.469</td>
<td>0.454</td>
<td>0.385</td>
<td>0.355</td>
<td>0.307</td>
<td>0.446</td>
<td>0.358</td>
<td>0.476</td>
<td>0.433</td>
<td>0.360</td>
<td>0.394</td>
<td>0.420</td>
</tr>
<tr>
<td>Common</td>
<td>0.484</td>
<td>0.408</td>
<td>0.446</td>
<td>0.373</td>
<td>0.400</td>
<td>0.536</td>
<td>0.452</td>
<td>0.478</td>
<td>0.465</td>
<td>0.444</td>
<td>0.386</td>
<td>0.359</td>
<td>0.314</td>
<td>0.448</td>
<td>0.357</td>
<td>0.480</td>
<td>0.439</td>
<td>0.354</td>
<td>0.405</td>
<td>0.426</td>
</tr>
<tr>
<td>SCL</td>
<td>0.467</td>
<td>0.395</td>
<td>0.453</td>
<td>0.391</td>
<td>0.415</td>
<td>0.531</td>
<td>0.434</td>
<td>0.484</td>
<td>0.461</td>
<td>0.461</td>
<td>0.387</td>
<td>0.377</td>
<td>0.333</td>
<td>0.449</td>
<td>0.361</td>
<td>0.488</td>
<td>0.439</td>
<td>0.342</td>
<td>0.401</td>
<td>0.431</td>
</tr>
<tr>
<td>ISF</td>
<td>0.474</td>
<td>0.434</td>
<td>0.455</td>
<td>0.446</td>
<td>0.405</td>
<td>0.537</td>
<td>0.454</td>
<td>0.484</td>
<td>0.504</td>
<td>0.460</td>
<td>0.386</td>
<td>0.399</td>
<td>0.423</td>
<td>0.357</td>
<td>0.375</td>
<td>0.355</td>
<td>0.338</td>
<td>0.282</td>
<td>0.373</td>
<td>0.316</td>
</tr>
</tbody>
</table>

26/10/2016
Conclusion

• A few NLP tasks
• Outperforming SCL