Cross Language Text Classification via Multi-View Subspace Learning

Yuhong Guo and Min Xiao

Dept. of Computer and Information Sciences
Temple University
Documents in different languages may share the same set of categories

- E.g., newsgroup dataset in English and French can cover the same set of categories

Standard monolingual classification methods

- Require sufficient number of labels in each language to train a monolingual classifier
- Expensive document annotation in each language

How about using labeled data in one language to help the classification in the other language via cross-language text classification?
Cross Language Text Classification

➤ Idea of cross-language text classification (CLTC):
  ❖ Exploit labeled data existing in language A to classify documents in language B
  ❖ Reduce expensive re-labeling process in language B

➤ Existing simple CLTC methods rely on machine translation:
  ❖ First translate documents from the source language A to the target language B, or vice versa
  ❖ Then apply standard monolingual classification
Cross Language Text Classification

- Two problems:
  - Feature distribution divergence between the original documents and the translated documents in each language
    - can be addressed by domain adaptation methods
  - Information loss and translation errors in machine translation process
    - can be alleviated by multi-view learning methods that exploit original data in both languages

- How about addressing these two types of problems simultaneously to gain more advantages?
Proposed Approach

- A Subspace Co-regularized Multi-view Learning Method
  - Translate the documents in each language into the other language using machine translation to form two parallel data matrices (two views): $X_1, X_2$
  - Exploit data from both views (languages) to alleviate the translation loss
  - Learn discriminative subspace representations of multi-view documents
    - capture the intrinsic structure of the data, bridge domain divergence
In each view, we project the data into low-dimensional subspace, and then use a linear prediction function. In the $i$th view, it is:

$$f_i(X_i^\ell) = X_i^\ell \Theta_i w_i + b_i$$

where $\Theta_i \in \mathbb{R}^{d_i \times m}$, $\Theta_i^T \Theta_i = I$ projects data into low-dimensional subspace
Two-view learning formulation:

\[
\min_{\{\Theta_i, w_i, b_i\}} \sum_{i=1}^{2} \| X_i^\ell \Theta_i w_i + b_i - y \|^2 + \alpha_i \| w_i \|^2 \\
+ \gamma \| X_1 \Theta_1 - X_2 \Theta_2 \|^2_F \\
\text{s. t.} \quad \Theta_1^T \Theta_1 = I, \quad \Theta_2^T \Theta_2 = I.
\]

Subspace co-regularization:
Distance of the two views on projected low-dimensional space
After solving the minimization over \( \{w_i, b_i\} \) for closed-form solutions, we obtain the following \textbf{orthogonal constrained} problem

\[
\min_{\Theta_1, \Theta_2} L(\Theta_1, \Theta_2) \quad \text{s. t.} \quad \Theta_1^\top \Theta_1 = I, \quad \Theta_2^\top \Theta_2 = I.
\]

\[
L(\Theta_1, \Theta_2) = \gamma \|X_1 \Theta_1 - X_2 \Theta_2\|_F^2 + 2y^\top H y - \sum_{i=1}^{2} z_i^\top \Theta_i (\Theta_i^\top M_i \Theta_i + \alpha_i I)^{-1} \Theta_i^\top z_i
\]
Optimization Algorithm

- Gradient descent with **curvilinear search**: 
  - requires no local projections 
  - always stays in feasible orthogonal region 
  - converges to local optimal solution
Curvilinear local search:

Given gradients:

\[ G_1 = \nabla_{\Theta_1} L(\Theta_1, \Theta_2), \quad G_2 = \nabla_{\Theta_2} L(\Theta_1, \Theta_2). \]

Compute skew symmetric matrices:

\[ F_1 = G_1 \Theta_1^T - \Theta_1 G_1^T, \quad F_2 = G_2 \Theta_2^T - \Theta_2 G_2^T \]

Such that

Curvilinear local search:

\[ Q_1(\tau) = \left( I + \frac{\tau}{2} F_1 \right)^{-1} \left( I - \frac{\tau}{2} F_1 \right) \Theta_1 \]
\[ Q_2(\tau) = \left( I + \frac{\tau}{2} F_2 \right)^{-1} \left( I - \frac{\tau}{2} F_2 \right) \Theta_2 \]

Note: \( Q_1(\tau)^T Q_1(\tau) = I \) and \( Q_2(\tau)^T Q_2(\tau) = I \)

Local descent path at \( \tau \geq 0 \)
Experiments

➢ Dataset

- A multilingual dataset: Reuters RCV1/RCV2
- 5 languages
  - English (E), French (F), German (G), Italian (I), Spanish (S)

➢ Comparison Approaches

- Baselines methods: TB, TSB
- Domain adaptation methods: EA++
- Multi-view co-classification method: MVMV, MVCC
- Proposed Method: SCMV
Table 1. Average classification accuracy results over 10 runs for 20 CLTC tasks.

<table>
<thead>
<tr>
<th>TASKS</th>
<th>TB</th>
<th>TSB</th>
<th>EA++</th>
<th>MVMV</th>
<th>MVCC</th>
<th>SCMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>E2F</td>
<td>78.60±0.80</td>
<td>79.24±0.51</td>
<td>79.52±0.47</td>
<td>81.13±0.46</td>
<td>83.20±0.38</td>
<td>86.10±0.42</td>
</tr>
<tr>
<td>E2G</td>
<td>75.65±0.67</td>
<td>75.01±0.51</td>
<td>75.25±0.46</td>
<td>80.37±0.76</td>
<td>81.62±0.54</td>
<td>83.51±0.74</td>
</tr>
<tr>
<td>E2I</td>
<td>79.80±0.69</td>
<td>76.39±0.98</td>
<td>76.48±1.02</td>
<td>80.01±0.69</td>
<td>83.75±0.64</td>
<td>84.87±0.51</td>
</tr>
<tr>
<td>E2S</td>
<td>84.54±1.52</td>
<td>85.24±1.01</td>
<td>85.43±1.03</td>
<td>86.30±0.69</td>
<td>89.98±0.42</td>
<td>92.26±0.34</td>
</tr>
<tr>
<td>F2E</td>
<td>77.04±0.92</td>
<td>80.32±0.47</td>
<td>80.60±0.48</td>
<td>81.15±0.44</td>
<td>82.51±0.36</td>
<td>83.86±0.35</td>
</tr>
<tr>
<td>F2G</td>
<td>76.41±0.92</td>
<td>76.32±0.62</td>
<td>76.68±0.49</td>
<td>79.66±0.91</td>
<td>81.84±0.76</td>
<td>83.16±0.70</td>
</tr>
<tr>
<td>F2I</td>
<td>78.32±0.82</td>
<td>77.02±0.78</td>
<td>78.87±0.75</td>
<td>79.53±0.63</td>
<td>82.98±0.47</td>
<td>83.25±0.43</td>
</tr>
<tr>
<td>F2S</td>
<td>84.77±1.05</td>
<td>86.24±0.71</td>
<td>86.90±0.69</td>
<td>87.53±0.68</td>
<td>90.96±0.44</td>
<td>92.81±0.25</td>
</tr>
<tr>
<td>G2E</td>
<td>77.04±0.88</td>
<td>78.57±0.37</td>
<td>78.42±0.36</td>
<td>78.68±0.68</td>
<td>80.52±0.50</td>
<td>82.52±0.47</td>
</tr>
<tr>
<td>G2F</td>
<td>75.93±0.70</td>
<td>77.08±0.51</td>
<td>77.22±0.42</td>
<td>77.99±0.61</td>
<td>80.57±0.48</td>
<td>83.55±0.36</td>
</tr>
<tr>
<td>G2I</td>
<td>79.88±0.77</td>
<td>78.54±1.05</td>
<td>78.61±0.99</td>
<td>78.07±0.78</td>
<td>81.85±0.54</td>
<td>84.20±0.51</td>
</tr>
<tr>
<td>G2S</td>
<td>85.82±0.91</td>
<td>86.22±0.55</td>
<td>86.61±0.57</td>
<td>84.73±0.62</td>
<td>89.24±0.37</td>
<td>90.67±0.61</td>
</tr>
<tr>
<td>I2E</td>
<td>76.98±0.74</td>
<td>76.76±0.42</td>
<td>77.80±0.40</td>
<td>78.86±0.61</td>
<td>80.45±0.47</td>
<td>81.34±0.48</td>
</tr>
<tr>
<td>I2F</td>
<td>76.88±0.94</td>
<td>78.10±0.35</td>
<td>78.61±0.47</td>
<td>78.11±0.65</td>
<td>80.58±0.60</td>
<td>81.73±0.42</td>
</tr>
<tr>
<td>I2G</td>
<td>76.79±0.57</td>
<td>76.56±0.55</td>
<td>77.66±0.48</td>
<td>79.69±0.61</td>
<td>80.50±0.53</td>
<td>84.76±0.35</td>
</tr>
<tr>
<td>I2S</td>
<td>85.36±1.42</td>
<td>87.68±0.50</td>
<td>88.63±0.51</td>
<td>89.42±0.56</td>
<td>90.66±0.33</td>
<td>94.15±0.44</td>
</tr>
<tr>
<td>S2E</td>
<td>74.35±0.94</td>
<td>74.73±0.63</td>
<td>74.83±0.69</td>
<td>77.89±0.54</td>
<td>79.45±0.58</td>
<td>80.50±0.44</td>
</tr>
<tr>
<td>S2F</td>
<td>75.89±1.10</td>
<td>77.48±0.58</td>
<td>77.62±0.57</td>
<td>77.93±0.62</td>
<td>82.82±0.22</td>
<td>84.86±0.33</td>
</tr>
<tr>
<td>S2G</td>
<td>75.88±0.44</td>
<td>74.28±0.40</td>
<td>74.31±0.34</td>
<td>77.91±0.56</td>
<td>80.90±0.44</td>
<td>81.12±0.53</td>
</tr>
<tr>
<td>S2I</td>
<td>79.36±0.84</td>
<td>79.72±0.69</td>
<td>80.54±0.75</td>
<td>82.46±0.65</td>
<td>87.18±0.46</td>
<td>88.59±0.47</td>
</tr>
</tbody>
</table>
Thanks!