Learning Domain Adaptation Classifiers from Multiple Distributed Sources

Doina Caragea
Computer Science
Kansas State University

SSC, Inference and Prediction from Distributed Data
May 31st, 2016
Acknowledgements

HongMin Li, KSU
Cornelia Caragea, UNT
Nic Herndon, KSU
Andrea Tapia, PSU

NSF
Beocat
K-INBRE
Twitter Mining to Aid Disaster Response
How Can One Identify Useful Content?

- Possible approaches:
  - Keyword search: noisy, non-relevant tweets retrieved
  - Manual selection: too time consuming
  - Supervised learning: a large number of labeled tweets needed
    [Imran et al., 2013; Ashktorab et al., 2014; Caragea et al., 2014]

- However:
  - Distributed data sources from previous disasters, collected and labeled by independent organizations, are available
  - Can we use such previous distributed data sources to identify useful information for a current disaster?
Domain Adaptation from Distributed Sources

Our approach [Li et al., 2015; Li et al., 2016]:

- Domain adaptation using labeled data from previous source disasters and unlabeled data from the current target disaster
Domain Adaptation

Challenge: distributions of source domains and target domain are generally different

- Need to identify knowledge that can be transferred from sources to target
- Need to identify approaches that can be used to transfer knowledge

Our approach to transferring knowledge:

- Naïve Bayes for Domain Adaptation [Herndon and Caragea, 2013; 2014] inspired from [Tan et al., 2009]
- An iterative approach of expectation maximization (EM)
Naïve Bayes Classifier

- Performs classification using the Bayes Theorem
- Makes the "naïve" assumption that features are independent given that class
- The class $c_x$ of a new instance $x = (x_1, \ldots, x_n)$ is

$$c_x = \arg \max_{c_k} P(c_k | x) = \arg \max_{c_k} \frac{P(x | c_k) \cdot P(c_k)}{P(x)} = \arg \max_{c_k} \prod_{i} P(x_i | c_k) \cdot P(c_k)$$

- Learning a Naïve Bayes classifier reduces to estimating class priors $P(c_k)$ and likelihoods $P(x_i | c_k)$ from the training data
Naïve Bayes Bernoulli Model for Tweets

Assumptions:

- words $w_i$ in a tweet data source are features
- a feature can take values 0/1

Probability estimates based on counts (a.k.a., sufficient statistics):

$$P(c_k) = \frac{N(c_k)}{N}$$

$$P(w_i = 0|c_k) = \frac{N(w_i = 0, c_k)}{N(c_k)}$$

$$P(w_i = 1|c_k) = \frac{N(w_i = 1, c_k)}{N(c_k)}$$

Easy to obtain in a distributed framework where there might be privacy concerns or concerns regarding the amount of data shipped through the network.
Single-Source Naïve Bayes for Domain Adaptation

1. Compute the prior and likelihood based on labeled source data, using only generalizable features from the source (where generalizable features are defined as source features that appear also in the target)

\[
P(c_k) = P_S(c_k) \\
P(w_i \mid c_k) = P_S(w_i \mid c_k)
\]
2. Assign labels to the unlabeled tweets from the target and identify tweets labeled with high confidence

\[ P(c_k \mid \text{tweet}) \propto P(c_k) \prod_{i=1}^{n} P(w_i \mid c_k) \]
Single-Source Naïve Bayes for Domain Adaptation

3. While labels assigned to unlabeled tweets change
   • **M-step**: compute prior and likelihood based on source and newly labeled target tweets

\[
P(c_k) = (1 - \lambda)P_S(c_k) + \lambda P_T(c_k)
\]
\[
P(w_i | c_k) = (1 - \lambda)P_S(w_i | c_k) + \lambda P_T(w_i | c_k)
\]
Single-Source Naïve Bayes for Domain Adaptation

3. While labels assigned to unlabeled tweets change
   - **E-step**: Assign labels to the unlabeled tweets from the target domain and identify new tweets labeled with high confidence.

\[
P(c_k \mid \text{tweet}) \propto P(c_k) \prod_{i=1}^{n} P(w_i \mid c_k)
\]
Single-Source Naïve Bayes for Domain Adaptation

4. Use the final classifier to assign labels to the test target tweets

\[ P(c_k \mid \text{tweet}) \propto P(c_k) \prod_{i=1}^{n} P(w_i \mid c_k) \]
Multi-Source Naïve Bayes for Domain Adaptation

- A straightforward generalization of the single-source Naïve Bayes for domain adaptation
- To estimate probabilities, combine weighted counts obtained from distributed sources

\[
P_S(c_k) = \frac{\sum_j^M \alpha_j \ast N_j(c_k)}{\sum_j^M \alpha_j \ast N_j}
\]

\[
P_S(w_i | c_k) = \frac{\sum_j^M \alpha_j \ast N_j(w_i, c_k)}{\sum_j^M \alpha_j \ast N_j(c_k)}
\]

where \( \alpha_i = \frac{1}{KL(D_T \| D_{S_j})} \) and

\[
KL(D_T \| D_{S_j}) = \sum_{w_i \in V} P_T(w_i) \log_2 \frac{P_T(w_i)}{P_{S_j}(w_i)}
\]

is the Kullback-Leibler divergence
Research Questions

• Does domain adaptation help?
  • Compare supervised Naïve Bayes classifiers with domain adaptation Naïve Bayes classifiers

• How does the performance vary with the amount of labeled data in the sources?
  • Vary the amount of source labeled data

• How does the performance vary with the number of sources?
  • Compare the results for domain adaptation with one source versus two sources versus three sources
### Data Sources

CrisisLexT6 [Olteanu et al., 2014]

<table>
<thead>
<tr>
<th>Disaster event</th>
<th>On-topic</th>
<th>Off-topic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012_Sandy_Hurricane</td>
<td>5,261</td>
<td>3,752</td>
<td>9,013</td>
</tr>
<tr>
<td>2013_Queensland_Floods</td>
<td>3,236</td>
<td>4,550</td>
<td>7,786</td>
</tr>
<tr>
<td>2013_Boston_Bombings</td>
<td>4,441</td>
<td>4,309</td>
<td>8,750</td>
</tr>
<tr>
<td>2013_Alfreda_Floods</td>
<td>3,497</td>
<td>4,714</td>
<td>8,211</td>
</tr>
</tbody>
</table>
Experimental Framework

• Use only target features to represent sources
• Perform 5-fold cross-validation over target and report average accuracy over the 5 folds (four target folds are used as unlabeled, while the fifth fold is used as test)
• Vary the number of instances in the sources (i.e., 500 instances per class, 1000 instances per class, etc.) - smaller datasets are subsets of the larger datasets
• Vary the number of sources (use 1, 2 or 3 sources at once)
Methods Compared

- NB: Supervised Naïve Bayes
- NB_SDA: Naïve Bayes single-source domain adaptation
- NB_MDA: Naïve Bayes multi-source domain adaptation
# Experimental Results: Accuracy

<table>
<thead>
<tr>
<th>Target</th>
<th>Source 1</th>
<th>Source 2</th>
<th>Source 3</th>
<th>Source 1,2</th>
<th>Source 2,3</th>
<th>Source 1,3</th>
<th>Source 1,2,3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2013_Alperta_Floods</strong></td>
<td>65.75</td>
<td>75.96</td>
<td>70.50</td>
<td>75.96</td>
<td>76.67</td>
<td>70.72</td>
<td>76.18</td>
</tr>
<tr>
<td><strong>2012_Sandy_Hurricane</strong></td>
<td>68.24</td>
<td>77.47</td>
<td>72.18</td>
<td>77.47</td>
<td>77.54</td>
<td>72.82</td>
<td>77.42</td>
</tr>
<tr>
<td><strong>2013_Queensland_Floods</strong></td>
<td>71.56</td>
<td>78.29</td>
<td>71.98</td>
<td>78.29</td>
<td>77.79</td>
<td>73.56</td>
<td>78.32</td>
</tr>
<tr>
<td><strong>2013_Boston_Bombing</strong></td>
<td>74.03</td>
<td>79.89</td>
<td>74.47</td>
<td>79.89</td>
<td>79.62</td>
<td>76.47</td>
<td>80.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Sources</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NB</strong></td>
<td>Source 1</td>
<td>65.75</td>
<td>68.24</td>
<td>71.56</td>
<td>74.03</td>
<td>71.06</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>75.90</td>
<td>76.17</td>
<td>76.40</td>
<td>77.41</td>
<td>78.87</td>
</tr>
<tr>
<td></td>
<td>Source 3</td>
<td>70.50</td>
<td>72.18</td>
<td>71.98</td>
<td>74.47</td>
<td>73.82</td>
</tr>
<tr>
<td></td>
<td>Sources 1,2</td>
<td>75.96</td>
<td>77.47</td>
<td>78.29</td>
<td>79.89</td>
<td>80.33</td>
</tr>
<tr>
<td></td>
<td>Sources 2,3</td>
<td>76.67</td>
<td>77.54</td>
<td>77.79</td>
<td>79.62</td>
<td>80.34</td>
</tr>
<tr>
<td></td>
<td>Sources 1,3</td>
<td>70.72</td>
<td>72.82</td>
<td>73.56</td>
<td>76.47</td>
<td>75.24</td>
</tr>
<tr>
<td></td>
<td>Sources 1,2,3</td>
<td>76.18</td>
<td>77.42</td>
<td>78.32</td>
<td>80.14</td>
<td>80.01</td>
</tr>
<tr>
<td><strong>NB_SDA (1)</strong></td>
<td>Source 1</td>
<td>76.76</td>
<td>79.3</td>
<td>79.92</td>
<td>83.97</td>
<td>82.57</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>81.83</td>
<td>82.35</td>
<td>82.95</td>
<td>85.18</td>
<td>86.01</td>
</tr>
<tr>
<td></td>
<td>Source 3</td>
<td>79.20</td>
<td>80.21</td>
<td>82.35</td>
<td>83.92</td>
<td>83.96</td>
</tr>
</tbody>
</table>
Experimental Results: Accuracy

<table>
<thead>
<tr>
<th>Target</th>
<th>Source 1</th>
<th>Source 2</th>
<th>Source 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013_Alabta_Floods</td>
<td>2012_Sandy_Hurricane</td>
<td>2013_Queensland_Floods</td>
<td>2013_Boston_Bombing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Sources</th>
<th>500</th>
<th>1000</th>
<th>2000</th>
<th>4000</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB_SDA (1)</td>
<td>Source 1</td>
<td>76.76</td>
<td>79.3</td>
<td>79.92</td>
<td>83.97</td>
<td>82.57</td>
</tr>
<tr>
<td></td>
<td>Source 2</td>
<td>81.83</td>
<td>82.35</td>
<td>82.95</td>
<td>85.18</td>
<td>86.01</td>
</tr>
<tr>
<td></td>
<td>Source 3</td>
<td>79.2</td>
<td>80.21</td>
<td>82.35</td>
<td>83.92</td>
<td>83.96</td>
</tr>
<tr>
<td>NB_MDA (2)</td>
<td>Sources 1,2</td>
<td>83.5</td>
<td>84.47</td>
<td>85.29</td>
<td>86.46</td>
<td>86.34</td>
</tr>
<tr>
<td></td>
<td>Sources 2,3</td>
<td>84.01</td>
<td>84.53</td>
<td>84.87</td>
<td>86.54</td>
<td>86.98</td>
</tr>
<tr>
<td>NB_MDA (3)</td>
<td>Sources 1,2,3</td>
<td>83.56</td>
<td>83.75</td>
<td>84.41</td>
<td>86.35</td>
<td>86.04</td>
</tr>
</tbody>
</table>
Conclusions

• Domain adaptation from multiple distributed sources helps improve the classification accuracy as compared to supervised learning

• Source data is useful even in small amounts (e.g., 500 instances), but larger amounts of source labeled data result in better accuracy

• Using more sources does not always mean better results
Future Work

- Experiment with more datasets, different application domains and classification problems
- Carefully tune the weights to establish a baseline for where we expect the results to be
- Improve the method for identifying importance weights for sources automatically
- Identify useful instances in each source and remove those that are not useful
- Evaluate the benefits of using a small amount of labeled target data