Explaining the Stars: Weighted Multiple-Instance Learning for Aspect-Based Sentiment Analysis

Written by Nikolaos Pappas and Andrei Popescu-Belis for EMNLP 2014
Presented by Xia Cui for NLP@UoL
Explaining the Stars?

• Sentiment
  • Classification(Positive/Negative)
  • Regression(Stars)

• Q: “Can we determine the ratings of each aspect automatically?”
  • Require large number of features designed to capture each aspect!

• A: (1) Word-level features
  • High-level linguistic features(e.g. PoS tagging, parsing)
  • Features that capture syntactic and semantic dependencies between words

(2) Linear model
  • e.g. Support Vector Regression with L2 norm, Lasso Regression with L1 norm
Problem

• Ignore the fact: sentences of text have diverse contributions to overall sentiment or a specific aspect

• Q: “To what extent does each part of a text contribute to the prediction of its overall sentiment or the rating of a particular aspect?”

• A: Multiple-Instance Regression (MIR)
  • Assume: not all the parts of text have the same distribution to predict rating
Multiple-Instance Regression (MIR)

- Multiple-Instance Learning (MIL) problem for real-valued output
- Each data point may be described by more than one vectors of values
- Learn internal structure of bags using clustering
MIR Definition

Input \( D = \{ (\{ b_{1j} \}_{n_1}, y_1), \ldots, (\{ b_{mj} \}_{n_m}, y_m) \} \)

Output \( \Phi = \arg \min_{\Phi} \left( \mathcal{L}(Y, X, \Phi) + \Omega(\Phi) \right) \)

- \( B \): set of bags
- \( m \): # bags
- \( n_i \): # instances (data points in \( B_i \), \( d \)-dimensional word vectors)
- \( Y \): labels
- \( X \): set of bag representations \( \leftarrow \) we need good assumption
- \( \Phi \): mapping (optimal regression hyperplane)
Previous Assumptions

- **Aggregate**
  - Each bag is represented as a single d-dimensional vector, which is the average of its instances.
  
  \[
  D_{agg} = \{(x_i, y_i) \mid i = 1, \ldots, m\}
  \]
  
  \[
  \hat{y}(B_i) = f(\text{mean}(\{b_{ij} \mid j = 1, \ldots, n_i\}))
  \]

- **Instance**
  - Each of the instances in a bag as separate examples, labelling each of them with the bag’s label. Train them all.
  
  \[
  D_{ins} = \{(b_{ij}, y_i) \mid j = 1, \ldots, n_i; i = 1, \ldots, m\}
  \]
  
  \[
  \hat{y}(B_i) = \text{mean}(\{f(b_{ij}) \mid j = 1, \ldots, n_i\})
  \]

- **Prime**
  - A single instance in each bag is responsible for its label. Train only one instance per bag.
  
  \[
  D_{pri} = \{(b_{i}^p, y_i) \mid i = 1, \ldots, m\}
  \]
  
  \[
  \hat{y}(B_i) = \text{mean}(\{f(b_{i}^p) \mid j = 1, \ldots, n_i\})
  \]
Proposed MIR Model

• Assumption: **Instance Relevance**
  • Each bag defines a bounded region of a hyperplane orthogonal to the y-axis
  • Find a regression hyperplane that passes through each bag $B_i$

\[ x_i = \sum_{j=1}^{n_i} \psi_{ij} b_{ij}, \psi_{ij} \geq 0 \text{ and } \sum_{j=1}^{n_i} \psi_{ij} = 1 \]

  • $\psi_{ij}$ is weight of $j_{th}$ instance of $i_{th}$ bag

• Modeling

\[ B_i = \{b_{ij}\}_{n_i}^d, \quad Y = \{y_i\}_N, y_i \in \mathbb{R}, \quad X = \{x_1, \ldots, x_m\}, \quad \psi_i = \{\psi_{ij}\}_{n_i}, \psi_{ij} \in [0, 1] \]

• Look for

\[ Y = f(X) = \Phi^T X \]
Modeling (Cont.)

• Loss Function

\[
\mathcal{L}(Y, X, \Psi, \Phi) = \|Y - \Phi^T X\|_2^2 \\
= \sum_{i=1}^{N} \left( y_i - \Phi^T \left( \sum_{j=1}^{n_i} \psi b_{ij} \right) \right)^2 \\
= \sum_{i=1}^{N} \left( y_i - \Phi^T (B_i \psi_i) \right)^2
\]

• Mapping

\[
\arg\min_{\psi_1, \ldots, \psi_m, \Phi} \sum_{i=1}^{m} \left( \Delta_i^2 + \epsilon_1 \|\psi_i\| \right) + \epsilon_2 \|\Phi\|^2 \\
\text{where } \Delta_i^2 = \left( y_i - \Phi^T (B_i \psi_i) \right)^2,
\]

• non-convex and difficult to optimize
Mapping (Cont.)

• Computed $f_1$: $\psi_1...\psi_m$, $f_2$: $\Phi$ (See next slide)

$$D_w = \{(b_{ij}, \psi_{ij}) \mid i = 1,...,m; j = 1,...,n_i\}$$

$$\arg \min\_{O} \sum\sum_{i=1}^{N} (\psi_{ij} - O^T b_{ij})^2 + e_3 \|O\|^2$$

• $O$: regression coefficients
• minimization can be solved by least squares

• Prediction
  • $\hat{\psi}_i = f_3(B_i) = \Omega^T B_i$
    $\hat{\psi}_i = \psi_i / \text{sum}(\psi_i)$ ← convert to $[0,1]$

  $$\hat{y} = f_2(B_i) = \Phi^T B_i \hat{\psi}_i$$
Learning

• Alternative Projection (AP)
  • Divide non-convex problem into two convex problems (i.e. $f_1, f_2$)
  • Find the intersection of convex sets

### Algorithm 1 APWeights($B, Y, \epsilon_1, \epsilon_2, \epsilon_3$)

1: Initialize($\psi_1, \ldots, \psi_N, \Phi, X$)
2: while not converged do
3:   for $B_i$ in $B$ do
4:     $\psi_i = cRLS(\Phi^T B_i, Y_i, \epsilon_1)$ # $f_1$ model
5:     $x_i = B_i \psi_i^T$
6:   end for
7:   $\Phi = RLS(X, Y, \epsilon_2)$ # $f_2$ model
8: end while
9: $\Omega = RLS(\{b_{ij} \forall i, j\}, \{\psi_{ij} \forall i, j\}, \epsilon_3)$ # $f_3$ model

• Complexity $T(h) = T_{ap}(h) + T_{f_3}(h)$
  $= O(m(n^2 + d^2)) + O(m \tilde{n} d^2)$
  $= O(m(n^2 + d^2 + \tilde{n} d^2))$,

( $\tilde{n}$: average size of bags)
Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Bags</th>
<th>Instances</th>
<th>Dimension</th>
<th>Aspect ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
<td>Count</td>
<td>Type</td>
<td>Count</td>
</tr>
<tr>
<td>BeerAdvocate</td>
<td></td>
<td>1,200</td>
<td></td>
<td>12,189</td>
</tr>
<tr>
<td>RateBeer (ES)</td>
<td></td>
<td>1,200</td>
<td></td>
<td>3,269</td>
</tr>
<tr>
<td>RateBeer (FR)</td>
<td></td>
<td>1,200</td>
<td>sentence</td>
<td>4,472</td>
</tr>
<tr>
<td>Audiobooks</td>
<td></td>
<td>1,200</td>
<td></td>
<td>4,886</td>
</tr>
<tr>
<td>Toys &amp; Games</td>
<td></td>
<td>1,200</td>
<td></td>
<td>6,463</td>
</tr>
<tr>
<td>TED comments</td>
<td></td>
<td>1,200</td>
<td>sentence</td>
<td>3,814</td>
</tr>
<tr>
<td>TED talks</td>
<td></td>
<td>1,200</td>
<td>comment</td>
<td>11,993</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Description of the seven datasets used for aspect, sentiment and emotion rating prediction.

Distributions of rating values per aspect rating class for the seven datasets.
Evaluation: Aspect Rating Performance

<table>
<thead>
<tr>
<th>Model \ Error</th>
<th>BeerAdvocate</th>
<th>RateBeer (ES)</th>
<th>RateBeer (FR)</th>
<th>Audiobooks</th>
<th>Toys &amp; Games</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MSE</td>
<td>MAE</td>
<td>MSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Aggregated (ℓ₁)</td>
<td>13.62</td>
<td>3.13</td>
<td>15.94</td>
<td>4.02</td>
<td>12.21</td>
</tr>
<tr>
<td>Aggregated (ℓ₂)</td>
<td>14.58</td>
<td>3.68</td>
<td>14.47</td>
<td>3.41</td>
<td>12.32</td>
</tr>
<tr>
<td>Instance (ℓ₁)</td>
<td>12.67</td>
<td>2.89</td>
<td>14.91</td>
<td>3.54</td>
<td>11.89</td>
</tr>
<tr>
<td>Instance (ℓ₂)</td>
<td>13.74</td>
<td>3.28</td>
<td>14.40</td>
<td>3.39</td>
<td>11.82</td>
</tr>
<tr>
<td>Prime (ℓ₁)</td>
<td>12.90</td>
<td>2.97</td>
<td>15.78</td>
<td>3.97</td>
<td>12.70</td>
</tr>
<tr>
<td>Prime (ℓ₂)</td>
<td>14.60</td>
<td>3.64</td>
<td>15.05</td>
<td>3.68</td>
<td>12.92</td>
</tr>
<tr>
<td>Clustering (ℓ₂)</td>
<td>13.95</td>
<td>3.26</td>
<td>15.06</td>
<td>3.64</td>
<td>12.23</td>
</tr>
<tr>
<td>APWeights (ℓ₂)</td>
<td>12.24</td>
<td>2.66</td>
<td>14.18</td>
<td>3.28</td>
<td>11.37</td>
</tr>
<tr>
<td>APW vs. SVR (%)</td>
<td>+16.0</td>
<td>+27.7</td>
<td>+2.0</td>
<td>+3.8</td>
<td>+7.6</td>
</tr>
<tr>
<td>APW vs. Lasso (%)</td>
<td>+10.1</td>
<td>+15.1</td>
<td>+11.0</td>
<td>+18.4</td>
<td>+6.8</td>
</tr>
<tr>
<td>APW vs. 2nd best (%)</td>
<td>+3.3</td>
<td>+7.8</td>
<td>+1.5</td>
<td>+3.3</td>
<td>+3.7</td>
</tr>
</tbody>
</table>
Evaluation (Cont.)
Evaluation: Sentiment and Emotion Prediction

<table>
<thead>
<tr>
<th>Model \ Error</th>
<th>SENT. LABELS TED comm.</th>
<th>EMO. LABELS TED talks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>AverageRating</td>
<td>19.47</td>
<td>5.05</td>
</tr>
<tr>
<td>Aggregated ($\ell_1$)</td>
<td>17.08</td>
<td>4.17</td>
</tr>
<tr>
<td>Aggregated ($\ell_2$)</td>
<td>16.88</td>
<td>4.47</td>
</tr>
<tr>
<td>Instance ($\ell_1$)</td>
<td>17.69</td>
<td>4.37</td>
</tr>
<tr>
<td>Instance ($\ell_2$)</td>
<td>16.93</td>
<td>4.24</td>
</tr>
<tr>
<td>Prime ($\ell_1$)</td>
<td>17.39</td>
<td>4.37</td>
</tr>
<tr>
<td>Prime ($\ell_2$)</td>
<td>18.03</td>
<td>4.91</td>
</tr>
<tr>
<td>Clustering ($\ell_2$)</td>
<td>17.64</td>
<td>4.34</td>
</tr>
<tr>
<td>APWeights ($\ell_2$)</td>
<td><strong>15.91</strong></td>
<td><strong>3.95</strong></td>
</tr>
<tr>
<td>$APW$ vs SVR (%)</td>
<td>+5.7</td>
<td>+11.5</td>
</tr>
<tr>
<td>$APW$ vs Lasso (%)</td>
<td>+6.8</td>
<td>+5.3</td>
</tr>
<tr>
<td>$APW$ vs 2nd (%)</td>
<td>+5.7</td>
<td>+5.3</td>
</tr>
</tbody>
</table>
### Examples of predicted sentiment

<table>
<thead>
<tr>
<th>Sentences per comment</th>
<th>$\hat{\psi}_k$</th>
<th>$\hat{\gamma}_k$</th>
<th>$\hat{\nu}_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Very brilliant and witty, as well as great improvisation.” “I enjoyed this one a lot.”</td>
<td>0.64</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>“That’s great idea, really like it!” “I can’t wait to try it, but first thing, I need a house with big windows, next year, maybe I can do that.”</td>
<td>0.56</td>
<td>4.2</td>
<td>4.0</td>
</tr>
<tr>
<td>“Unfortunately countries are not led by gifted children.” “They are either dictated by the most extreme personalities who crave nothing but power or managed by politicians who are voted in by a far from gifted population.”</td>
<td>0.48</td>
<td>2.4</td>
<td>2.0</td>
</tr>
<tr>
<td>“I am very disappointed by this, smug, cliched and missing so much information as to be almost (...)” “No mention of ship transport lets say 50% of all material transport, no mention of rail transport, (...)” “I am sorry to be so negative, this just sounds like a sales pitch that he has given too many times (...)”</td>
<td>0.43</td>
<td>1.8</td>
<td>1.0</td>
</tr>
</tbody>
</table>

### Top comments

<table>
<thead>
<tr>
<th>Class</th>
<th>Top comment per talk (according to weights $\hat{\psi}_k$)</th>
<th>$\hat{\psi}_k$ distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>inspiring</td>
<td>“It seems to me that the idea worth spreading of this TED Talk is inspiring and key for a full life. ‘No-one else is the authority on your potential. You’re the only person that decides how far you go and what you’re capable of.’ It seems to me that teens actually think that. As a child one is all knowing and all capable. How did we get to the (...)”</td>
<td><img src="image1" alt="Graph" /></td>
</tr>
<tr>
<td>beautiful</td>
<td>“The beauty of the nature. It would be more interesting just integrates his thought and idea into a mobile device, like a mobile, so we can just turn on the nature gallery in any time. The paintings don’t look incidental but genuinely thought out, random perhaps, but with a clear grand design behind the randomness. Drawing is an art where it doesn’t (...)”</td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td>funny</td>
<td>“Funny story, but not as funny as a good ‘knock-knock’ joke. My favorite knock-knock joke of all time is Cheech &amp; Chong’s ‘Dave’s Not Here’ gag from the early 1970s. I’m still waiting for someone to top it after all these years. [Knock, knock] ‘Who is it?’ the voice of an obviously stoned man answers from the other side of a door, (...)”</td>
<td><img src="image3" alt="Graph" /></td>
</tr>
<tr>
<td>courageous</td>
<td>“I was a soldier in Iraq and part of the unit represented in this documentary. I would question anyone that told you we went over there to kill Iraqi people. I spent the better part of my time in Iraq protecting the Iraqi people from insurgents who came from countries outside of Iraq to kill Iraqi people. We protected families men, women, and (...)”</td>
<td><img src="image4" alt="Graph" /></td>
</tr>
</tbody>
</table>

### Top words

- inspiring
- beautiful
- funny
- courageous
- help
- government
- song
- music
- art
- and
- family
Conclusion and Future Work

• Contribution
  • Complexity competitive
  • Assign instance relevance on labelled bags, and predict them on unseen bags
  • Better performance

• Future
  • Test on sentence-level sentiment classification based on document-level
  • Experiment with other model settings
  • Investigate new method to estimate instance weights at prediction time
  • Evaluate impact of assigned weights on sentence ranking, segmentation or summarization

29/06/2016