Large Margin Taxonomy Embedding with an Application to Document Categorization

K. Weinberger and O. Chapelle
NIPS 2008

presented by J. Silva, Duke University
Problem

- Multi-class classification
  - Application: document categorization
- The classes (topics) follow a taxonomy (e.g., a hierarchy)
- Misclassification errors are not all the same. Examples:
  - It is worse to misclassify a male pedestrian as a traffic light than as a female pedestrian
  - ...or to misclassify a medical journal on heart attack as a publication on athlete’s foot than on coronary disease
- The proposed approach is cost-sensitive, and aims to move beyond hierarchical representations to discover a continuous latent semantic space
Multi-class classification of documents based on a taxonomy of topics

- $\vec{x}_i$ is the $i$-th document
- $\vec{p}_\alpha$ is the prototype for the $\alpha$-th class
- $P$, $W$ are mappings to latent semantic space

Note: All figures adapted from the original paper

Large Margin Taxonomy Embedding with an Application to Document Categorization
May 13, 2011 3 / 16
Contributions

- A supervised regression algorithm called *taxem* (taxonomy embedding) is presented.
- The algorithm learns the regression for the documents and the placement of the topic prototypes in a single optimization.
- The regression is done by solving a convex semi-definite programming (SDP) problem.
- In this case, the SDP admits a particular form that can be solved efficiently for large datasets.
Outline of the presentation

- Notation
- Two-step method
  - Topic embedding
  - Document regression
- One-step combined large margin optimization
- Results on the OHSUMED medical journal database
- Discussion
Documents $\vec{x}_1, \ldots, \vec{x}_n \in \mathcal{X}$ of dimensionality $d$
  ▶ Can be, e.g., bag-of-word indicators or tf-idf scores

$y_1, \ldots, y_n \in \{1, \ldots, c\}$ are topic labels in some taxonomy $\mathcal{T}$

Indices
  ▶ $i, j \in \{1, \ldots, n\}$ for documents
  ▶ $\alpha, \beta \in \{1, \ldots, c\}$ for classes

The taxonomy gives rise to a cost matrix $C \in \mathbb{R}^{c \times c}$, where $C_{\alpha\beta} \geq 0$ is the cost of misclassifying class $\alpha$ as $\beta$ and $C_{\alpha\alpha} = 0$

We wish to represent
  ▶ each topic $\alpha$ as a prototype $\vec{p}_{\alpha} \in \mathcal{F}$
  ▶ each document $\vec{x}_i$ as a low-dimensional vector $\vec{z}_i \in \mathcal{F}$

We assume $C$ is given
Two-step approach: embedding topic prototypes

- Find prototypes $\bar{p}_1, \ldots, \bar{p}_c \in \mathcal{F}$ based on $\mathbf{C}$
- Define $\mathbf{P} = [\bar{p}_1 \cdots \bar{p}_c] \in \mathbb{R}^{c \times c}$
  (note: it is assumed throughout the paper that $\mathcal{F} = \mathbb{R}^c$)
- How to derive $\mathbf{P}$ from $\mathbf{C}$?
  - Simplest: ignore $\mathbf{C}$ and set $\mathbf{P} = \mathbf{I}_{c \times c}$; all topics in the corners of a $(c-1)$-dimensional simplex, denoted $\mathbf{P}_I$
  - Better: solve

$$\mathbf{P}_{mds} = \arg \min_{\mathbf{P}} \sum_{\alpha,\beta=1}^{c} (\|\bar{p}_\alpha - \bar{p}_\beta\|_2^2 - C_{\alpha\beta})^2$$

where $mds$ means *metric dimensional scaling*, as in ISOMAP

- The solution is $\mathbf{P}_{mds} = \sqrt{\Lambda} \mathbf{V}$, obtained from the decomposition $\bar{\mathbf{C}} = \mathbf{V} \Lambda \mathbf{V}^\top$ where $\bar{\mathbf{C}} = -\frac{1}{2} \mathbf{H} \mathbf{C} \mathbf{H}$ and $\mathbf{H} = \mathbf{I} - \frac{1}{c} \mathbf{1} \mathbf{1}^\top$
Two-step approach: document regression

- Assume we have $P$
- Find mapping $W : \mathcal{X} \rightarrow \mathcal{F}$ so that $\vec{x}_i$ with label $y_i$ is placed near $\vec{p}_{y_i}$
- Solve the linear ridge regression

$$W = \arg \min_W \sum_{i=1}^{n} (\|\vec{p}_{y_i} - W\vec{x}_i\|_2^2 + \lambda\|W\|_F^2)$$

- The solution has the closed form $PJX^\top (XX^\top + \lambda I)^{-1}$ where $X = [\vec{x}_1 \cdots \vec{x}_n]$ and $J \in \{0, 1\}^{c \times n}$, with $J_{\alpha i} = 1$ iff $y_i = \alpha$
Inference and performance measure

- **Inference (for new documents)**
  - Given a new document $\vec{x}_t$, first map it into $F$ and then estimate its label via the nearest-neighbor rule
    \[
    \hat{y}_t = \arg\min_{\alpha} \|\vec{p}_\alpha - W\vec{x}_t\|_2^2
    \]

- **Performance measure**
  - For a given set of labeled documents $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$, the quality of the regression is assessed via the averaged cost-sensitive misclassification loss
    \[
    \mathcal{E} = \frac{1}{n} \sum_{i=1}^{n} C_{y_i\hat{y}_i}
    \]
**One-step method**

- Learning the prototypes independently of the data is not optimal
- Untangle the mutual dependence between $\mathbf{W}$ and $\mathbf{P}$ ("chicken-and-egg" problem)
  - Define
    
    $$
    \mathbf{A} = \mathbf{JX}^\top (\mathbf{XX}^\top + \lambda \mathbf{I})^{-1}
    $$

  - We have $\mathbf{W} = \mathbf{PA}$; $\mathbf{A}$ is independent of $\mathbf{P}$, and can be pre-computed

- Now find $\mathbf{P}$
  - Let $\tilde{\mathbf{x}}_i' = \mathbf{A} \tilde{\mathbf{x}}_i$ and $\tilde{e}_\alpha = [0 \cdots 1 \cdots 0]^\top$ with 1 in the $\alpha$-th position
  - Rewrite $\tilde{p}_\alpha = \mathbf{P} \tilde{e}_\alpha$ and $\tilde{z}_i = \mathbf{P} \tilde{x}_i'$
  - We can’t optimize $\mathcal{E}$ w.r.t. $\mathbf{P}$ directly (the objective is non-continuous and non-differentiable)
One-step method (cont.)

Surrogate loss function

Minimize $\sum_{i,\alpha} \xi_{i\alpha}$ s.t.

- $\|P(\tilde{e}_{y_i} - \tilde{x}_i')\|_2^2 + C_{y_i\alpha} \leq \|P(\tilde{e}_\alpha - \tilde{x}_i')\|_2^2 + \xi_{i\alpha}$
- $\xi_{i\alpha} \geq 0$

- This enforces a large-margin condition, so that prototypes which would incur larger misclassification loss are farther away
- The surrogate loss is an upper bound on $\epsilon$
One-step method (cont.)

Figure 2: The schematic layout of the large-margin embedding of the taxonomy and the documents. As a first step, we represent topic $\alpha$ as the vector $\vec{e}_\alpha$ and document $\vec{x}_i$ as $\vec{x}'_i = A \vec{x}_i$. We then learn the matrix $P$ whose columns are the prototypes $\vec{p}_\alpha = P \vec{e}_\alpha$ and which defines the final transformation of the documents $\vec{z}'_i = P \vec{x}'_i$. This final transformation is learned such that the correct prototype $\vec{p}_{y_i}$ is closer to $\vec{z}'_i$ than any other prototype $\vec{p}_\alpha$ by a large margin.
One-step method: upper bound on $E$

**Theorem 1** Given a prototype matrix $P$, the training error (5) is bounded above by $\frac{1}{n} \sum_{i,\alpha} \xi_{i\alpha}$.

**Proof:** First, note that we can rewrite the assignment of the closest prototype (4) as $\hat{y}_i = \text{argmin}_\alpha \|P(\bar{c}_\alpha - \bar{x}_i')\|^2$. It follows that $\|P(\bar{c}_{y_i} - \bar{x}_i')\|^2 - \|P(\bar{c}_{\hat{y}_i} - \bar{x}_i')\|^2 \geq 0$ for all $i$ (with equality when $\hat{y}_i = y_i$). We therefore obtain:

$$\xi_{i\hat{y}_i} = \|P(\bar{c}_{y_i} - \bar{x}_i')\|^2 + C_{y_i\hat{y}_i} - \|P(\bar{c}_{\hat{y}_i} - \bar{x}_i')\|^2 \geq C_{y_i\hat{y}_i}. \quad (9)$$

The result follows immediately from (9) and that $\xi_{i\alpha} \geq 0$:

$$\sum_{i,\alpha} \xi_{i\alpha} \geq \sum_{i} \xi_{i\hat{y}_i} \geq \sum_{i} C_{y_i\hat{y}_i}. \quad (10)$$
The above optimization is not convex, due to quadratic constraints

Make the problem invariant to rotations by defining $Q = P^T P$ and writing distances w.r.t. $Q$

$$\|P(\bar{e}_\alpha - \bar{x}_i')\|_2^2 = (\bar{e}_\alpha - \bar{x}_i')^T Q (\bar{e}_\alpha - \bar{x}_i') = \|\bar{e}_\alpha - \bar{x}_i'\|_Q^2$$

$\mu$ is a regularization parameter

Minimize $(1 - \mu) \sum_{i, \alpha} \xi_{i\alpha} + \mu \|Q - \bar{C}\|_F^2$ subject to:

1. $\|\bar{e}_{y_i} - \bar{x}_i'\|_Q^2 + C_{y_i\alpha} \leq \|\bar{e}_\alpha - \bar{x}_i'\|_Q^2 + \xi_{i\alpha}$
2. $\xi_{i\alpha} \geq 0$
3. $Q \succeq 0$
Results

<table>
<thead>
<tr>
<th>Top category</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td># samples n</td>
<td>7544</td>
<td>4772</td>
<td>4858</td>
<td>2701</td>
<td>7300</td>
<td>1961</td>
<td>8694</td>
<td>8155</td>
</tr>
<tr>
<td># topics c</td>
<td>424</td>
<td>160</td>
<td>453</td>
<td>339</td>
<td>457</td>
<td>151</td>
<td>425</td>
<td>150</td>
</tr>
<tr>
<td># nodes</td>
<td>519</td>
<td>312</td>
<td>610</td>
<td>608</td>
<td>559</td>
<td>218</td>
<td>533</td>
<td>170</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the different OHSUMED problems. Note that not all nodes are populated and that we pruned all strictly un-populated subtrees.
Results

<table>
<thead>
<tr>
<th>data</th>
<th>SVM l/all</th>
<th>MCSVM</th>
<th>SVM cost</th>
<th>SVM tax</th>
<th>P_l-taxem</th>
<th>P_mds-taxem</th>
<th>LM-taxem</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.17</td>
<td>2.13</td>
<td>2.11</td>
<td>1.96</td>
<td>2.11</td>
<td>2.33</td>
<td>1.95</td>
</tr>
<tr>
<td>B</td>
<td>1.50</td>
<td>1.38</td>
<td>1.64</td>
<td>1.52</td>
<td>1.57</td>
<td>1.99</td>
<td>1.39</td>
</tr>
<tr>
<td>C</td>
<td>2.41</td>
<td>2.32</td>
<td>2.25</td>
<td>2.25</td>
<td>2.30</td>
<td>2.61</td>
<td>2.16</td>
</tr>
<tr>
<td>D</td>
<td>3.10</td>
<td>2.76</td>
<td>2.92</td>
<td>2.82</td>
<td>2.82</td>
<td>3.05</td>
<td>2.66</td>
</tr>
<tr>
<td>E</td>
<td>3.44</td>
<td>3.42</td>
<td>3.26</td>
<td>3.25</td>
<td>3.45</td>
<td>3.05</td>
<td>3.05</td>
</tr>
<tr>
<td>F</td>
<td>2.59</td>
<td>2.65</td>
<td>2.66</td>
<td>2.69</td>
<td>2.63</td>
<td>2.77</td>
<td>2.51</td>
</tr>
<tr>
<td>G</td>
<td>3.98</td>
<td>4.12</td>
<td>3.89</td>
<td>3.82</td>
<td>4.10</td>
<td>3.63</td>
<td>3.59</td>
</tr>
<tr>
<td>H</td>
<td>2.42</td>
<td>2.48</td>
<td>2.40</td>
<td>2.32</td>
<td>2.45</td>
<td>2.24</td>
<td>2.17</td>
</tr>
<tr>
<td>all</td>
<td>2.78</td>
<td>2.77</td>
<td>2.77</td>
<td>2.65</td>
<td>2.79</td>
<td>2.73</td>
<td>2.50</td>
</tr>
</tbody>
</table>

Table 2: The cost-sensitive test error results on various ohsumed classification data sets. The algorithms are from left to right: one vs. all SVM, MCSVM [6], cost-sensitive MCSVM, Hierarchical SVM [4], simplex regression, mds regression, large-margin taxem. The best results (up to statistical significance) are highlighted in bold. The taxem algorithm obtains the lowest overall loss and the lowest individual loss on each data set except B.