Two-way Multimodal Online Matrix Factorization

JORGE A. VANEGAS

MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
Outline

1. Introduction
2. Two-way Multimodal Online Matrix Factorization for Multi-label Annotation
3. Semi-supervised Dimensionality Reduction via Multimodal Matrix Factorization
Semantic gap

Synonymy and polysemy

Text

Semantic similarity

Visual similarity

Images
Dimensionality reduction

Eliminate the **redundancy** and the **noise** present in the manifold structure of the original high dimensional feature representation.

Tackles the **curse of dimensionality** by compressing the representation in a more expressive reduced set of variables.
Semantic representation via matrix factorization

- Latent Semantic Analysis (LSA)  
  [Dumais et al. 2004]

- Nonnegative Matrix Factorization (NMF)  
  [Lee et al. 1999]
Two-way Multimodal Online Matrix Factorization for Multi-label Annotation (ICPRAM)

Jorge A. Vanegas

MindLAB Research Group - Universidad Nacional de Colombia

November 24, 2015
The multi-label annotation problem arises in situations such as object recognition in images where we want to automatically find the objects present in a given image. The solution consists in learning a classification model able to assign one or many labels to a particular sample.
Two-way Multimodal Matrix Factorization

\[
\begin{align*}
X & \xrightarrow{f} H & \xrightarrow{g} X \\
\end{align*}
\]

\[
f : \mathbb{R}^n \rightarrow \mathbb{R}^r,
\]

\[
g : \mathbb{R}^r \rightarrow \mathbb{R}^n
\]

where \( n \gg r \)

Reconstruction of the original feature representation:

\[
X_v \approx W'_v W_v X_v \tag{1}
\]

Reconstruction of the original label representation:

\[
X_t \approx W'_t W_t X_t \tag{2}
\]

Jorge A. Vanegas

Two-way Multimodal Online Matrix Factorization for Multi-label
Two-way Multimodal Matrix Factorization

\[ X_t = W'_t W_v X_v \]
Optimization problem

\[
L \left( X_v, X_t, W_v, W_v', W_t, W_t' \right) = \\
\alpha \left\| X_v - W_v' W_v X_v \right\|_F^2 + \\
+ (1 - \alpha) \left\| X_t - W_t' W_t X_t \right\|_F^2 \\
+ \delta \left\| X_t - W_t' W_v X_v \right\|_F^2 \\
+ \beta \left( \left\| W_v \right\|_F^2 + \left\| W_v' \right\|_F^2 + \left\| W_t \right\|_F^2 + \left\| W_t' \right\|_F^2 \right) \\
\] (3)

\( \alpha \) controls the relative importance between the reconstruction of the instance representation and the label representation.

\( \delta \) controls the relative importance of the mapping between instance features and label information.

\( \beta \) controls the relative importance of the regularization terms, which penalizes large values for the Frobenius norm of the transformation matrices.
Online learning

Loss function:

\[ Q(z, w) = \ell(f_w(x), y) \]

Gradient descent:

\[ w_{t+1} = w_t - \gamma \frac{1}{n} \sum_{i=1}^{n} \nabla_w Q(z_i, w_t), \]

Stochastic gradient descent:

\[ w_{t+1} = w_t - \gamma_t \nabla_w Q(z_t, w_t). \]
Online learning

• The learning rate $\gamma$ can be either constant or gradually decaying
• "Generally" move in the direction of the global minimum, but not always
• Never actually converges like batch gradient descent does, but ends up wandering around some region close to the global minimum. In practice, this isn't a problem

![Diagram showing learning curves and convergence]

Jorge A. Vanegas
Two-way Multimodal Online Matrix Factorization for Multi-label
Two-way multimodal online matrix factorization algorithm

**input** $r$: latent space size, $\gamma_0$: initial step size, *epochs*: number of epochs, $X_v \in \mathbb{R}^{n \times l}$, $X_t \in \mathbb{R}^{m \times l}$, $\alpha$, $\delta$, $\beta$

**Random initialization of transformation matrices:**

- $W_v^{(0)} = \text{random\_matrix}(r, n)$
- $W_v^{(0)} = \text{random\_matrix}(n, r)$
- $W_t^{(0)} = \text{random\_matrix}(r, m)$
- $W_t^{(0)} = \text{random\_matrix}(m, r)$

**for** $i = 1$ **to** *epochs* **do**

**for** $j = 1$ **to** $l$ **do**

**for** $\tau = i \times j$

- $x_v^{(\tau)}, x_t^{(\tau)} \leftarrow \text{sample\_without\_replacement}(X_v, X_t)$

**Compute gradients:**

- $g_{W_v}^{(\tau)} = \nabla_{W_v} L \left( x_v^{(\tau)}, x_t^{(\tau)}, W_v^{(\tau)}, W_v^{(\tau)}, W_t^{(\tau)}, W_t^{(\tau)} \right)$
- $g_{W_v}^{(\tau)} = \nabla_{W_v} L \left( x_v^{(\tau)}, x_t^{(\tau)}, W_v^{(\tau)}, W_v^{(\tau)}, W_t^{(\tau)}, W_t^{(\tau)} \right)$
- $g_{W_t}^{(\tau)} = \nabla_{W_t} L \left( x_v^{(\tau)}, x_t^{(\tau)}, W_v^{(\tau)}, W_v^{(\tau)}, W_t^{(\tau)}, W_t^{(\tau)} \right)$
- $g_{W_t}^{(\tau)} = \nabla_{W_t} L \left( x_v^{(\tau)}, x_t^{(\tau)}, W_v^{(\tau)}, W_v^{(\tau)}, W_t^{(\tau)}, W_t^{(\tau)} \right)$

...
Two-way multimodal online matrix factorization algorithm

... Update term calculation using momentum:

\[
\begin{align*}
\triangle W_{\nu}^{(\tau)} &= -\gamma^{(\tau)} g_{W_{\nu}}^{(\tau)} + p \triangle W_{\nu}^{(\tau-1)} \\
\triangle W_{\nu}^{(\tau)} &= -\gamma^{(\tau)} g_{W_{\nu}}^{(\tau)} + p \triangle W_{\nu}^{(\tau-1)} \\
\triangle W_{\nu}^{(\tau)} &= -\gamma^{(\tau)} g_{W_{\nu}}^{(\tau)} + p \triangle W_{\nu}^{(\tau-1)} \\
\triangle W_{\nu}^{(\tau)} &= -\gamma^{(\tau)} g_{W_{\nu}}^{(\tau)} + p \triangle W_{\nu}^{(\tau-1)} \\
\end{align*}
\]

Update transformation matrices:

\[
\begin{align*}
W_{\nu}^{(\tau+1)} &= W_{\nu}^{(\tau)} + \triangle W_{\nu}^{(\tau)} \\
W_{\nu}^{(\tau+1)} &= W_{\nu}^{(\tau)} + \triangle W_{\nu}^{(\tau)} \\
W_{\nu}^{(\tau+1)} &= W_{\nu}^{(\tau)} + \triangle W_{\nu}^{(\tau)} \\
W_{\nu}^{(\tau+1)} &= W_{\nu}^{(\tau)} + \triangle W_{\nu}^{(\tau)} \\
\end{align*}
\]

end for

end for

return $W_{\nu}^{(N)}, W_{\nu}^{(N)}, W_{\nu}^{(N)}, W_{\nu}^{(N)}$
Prediction

\[ \tilde{x}_t = W_t' W_v x_v \]
The transformation of the input features generates an $m$-dimensional vector with a smoothed label representation.

- The final decision to assign a label would be taken by defining a threshold.
- Assign 1 to the $j^{th}$ label if $\tilde{x}_{t,j} \geq \text{threshold}$, or we can assign 1 to the top-$k$ labels with the highest values in the vector.
Implementation details

- **Pylearn2 library**
- **Theano**

---

1. [http://deeplearning.net/software/pylearn2/](http://deeplearning.net/software/pylearn2/)
2. [http://deeplearning.net/software/theano/](http://deeplearning.net/software/theano/)
Experiments and Results

- 80% of the images for training
- the remaining 20% for test
- Were compared against 8 MLLSE\(^3\) algorithms:
  - OVA: One-versus-all
  - CCA: Canonical Correlation Analysis
  - CS: Compressed Sensing
  - PLST: Principal Label Space Transform
  - MME: Multilabel max-margin embedding
  - ANMF, MNMF, OMMF

\(^3\) multi-label latent space embedding
Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Corel5k</th>
<th>Bibtex</th>
<th>MediaMill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labels</td>
<td>374</td>
<td>159</td>
<td>101</td>
</tr>
<tr>
<td>Features</td>
<td>500</td>
<td>1,836</td>
<td>120</td>
</tr>
<tr>
<td>Label cardinality</td>
<td>3,522</td>
<td>2,402</td>
<td>4.376</td>
</tr>
<tr>
<td>Examples</td>
<td>5,000</td>
<td>7,395</td>
<td>43,907</td>
</tr>
</tbody>
</table>

The method was evaluated on 3 standard multilabel datasets distributed by the mulan framework authors (http://mulan.sourceforge.net/datasets.html)
## F-Measure for each method

### Performance of each method in terms of f-measure

<table>
<thead>
<tr>
<th>Method</th>
<th>Corel5k</th>
<th>Bibtex</th>
<th>MediaMill</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVA</td>
<td>0.112</td>
<td>0.372</td>
<td>—</td>
</tr>
<tr>
<td>CCA</td>
<td>0.150</td>
<td>0.404</td>
<td>—</td>
</tr>
<tr>
<td>CS</td>
<td>0.086 (50)</td>
<td>0.332 (50)</td>
<td>—</td>
</tr>
<tr>
<td>PLST</td>
<td>0.074 (50)</td>
<td>0.283 (50)</td>
<td>—</td>
</tr>
<tr>
<td>MME</td>
<td>0.178 (50)</td>
<td>0.403 (50)</td>
<td>0.199 (350)</td>
</tr>
<tr>
<td>ANMF</td>
<td>0.210 (30)</td>
<td>0.297 (140)</td>
<td>0.496 (350)</td>
</tr>
<tr>
<td>MNMF</td>
<td>0.240 (35)</td>
<td>0.376 (140)</td>
<td>0.510 (350)</td>
</tr>
<tr>
<td>OMMF</td>
<td>0.263 (40)</td>
<td><strong>0.436 (140)</strong></td>
<td>0.503 (350)</td>
</tr>
<tr>
<td>Our Method</td>
<td><strong>0.283 (100)</strong></td>
<td>0.422 (300)</td>
<td><strong>0.540 (300)</strong></td>
</tr>
</tbody>
</table>

Sebastian Otálora-Montenegro et al. [1]

Jorge A. Vanegas

Two-way Multimodal Online Matrix Factorization for Multi-label
We presented a novel multi-label annotation method which learns a mapping between the original sample representation and labels by finding a common semantic representation. We propose a model that finds a mapping from the sample representation space to a semantic space, and simultaneously finds a back-projection from the semantic space to the original space. This method is formulated as an online learning algorithm allowing to deal with large collections. One important limitation is that the method assumes linear dependencies between the data modalities.
Sebastian Otálora-Montenegro, Santiago A. Pérez-Rubiano, and Fabio A. González.
Online matrix factorization for space embedding multilabel annotation.

Sunho Park and Seungjin Choi.
Max-margin embedding for multi-label learning.
Semi-supervised Dimensionality Reduction via Multimodal Matrix Factorization

VIVIANA BELTRÁN, JORGE A. VANE GAS, FABIO A. GONZÁLE Z

MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
## Motivation

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Proposed Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>High dimensional data</td>
<td>Low dimensional embedding representation</td>
</tr>
<tr>
<td>Reduced number of labeled data</td>
<td>Semi–supervised learning via matrix factorization</td>
</tr>
<tr>
<td>Huge amount of unstructured data:</td>
<td>• Stochastic gradient descent</td>
</tr>
<tr>
<td>• Massive unlabeled examples available</td>
<td>• GPU implementation</td>
</tr>
</tbody>
</table>
Semi-supervised learning

Manifold learning for classification
Semi-supervised learning

Annotated instances are used to maximize the discrimination between classes, but also, non-annotated instances can be exploited to estimate the intrinsic manifold structure of the data.
Model

\[ f : \mathbb{R}^n \rightarrow \mathbb{R}^r, \]

\[ g : \mathbb{R}^r \rightarrow \mathbb{R}^n \]

\[ n \gg r \]

\[ X_x \approx W'_x W_x X_x \]

\[ X_t \approx W'_t W_t X_t \]
Model

\[ L = \alpha \sum_{i=1}^{k} \left\| x_i - W'_x W_x x_i \right\|_F^2 + (1 - \alpha) \sum_{i=1}^{l} \left\| t_i - W'_t W_t t_i \right\|_F^2 \]

\[ \quad + \delta \sum_{i=1}^{l} \left\| t_i - W'_t W_x x_i \right\|_F^2 + \beta \left( \left\| W_v \right\|_F^2 + \left\| W'_v \right\|_F^2 + \left\| W_t \right\|_F^2 + \left\| W'_t \right\|_F^2 \right) \]

\[ X_i \quad \text{feature vector of the } i-th \text{ instance in the data collection } X \]
\[ t_i \quad \text{binary label vector of the } i-th \text{ instance} \]
\[ k \quad \text{instances for training} \]
\[ l \quad \text{labeled instances} \]

\[ k \gg l \]
Experiments

Experimental Setup:

- Parameter exploration by using 5-fold cross validation.
- Average classification accuracies for 10 runs evaluated by using 1-KNN setup similar as in [Zhao et al. 2014].
- Linear supervised, semi-supervised and unsupervised dimensionality reduction methods as baselines.
Datasets

- MNIST
- USPS
- LETTERS
- COVTYPE
## Datasets

### Datasets partitions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Original dataset partitions</th>
<th>Low-scale partitions</th>
<th>Large-scale evaluation</th>
<th>#Dim</th>
<th>#Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td>Train</td>
<td>Test</td>
<td>Train</td>
</tr>
<tr>
<td>Covtype</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MNIST</td>
<td>60000</td>
<td>10000</td>
<td>8000</td>
<td>8000</td>
<td>60000</td>
</tr>
<tr>
<td>Letters</td>
<td>20000</td>
<td></td>
<td>8000</td>
<td>8000</td>
<td></td>
</tr>
<tr>
<td>USPS</td>
<td>4649</td>
<td>4649</td>
<td>4649</td>
<td>4649</td>
<td></td>
</tr>
</tbody>
</table>
## Classification accuracy for different percentages of annotated instances

<table>
<thead>
<tr>
<th>METHOD</th>
<th>STWOMF  r=C</th>
<th>STWOMF r=C+10</th>
<th>SDA</th>
<th>LDA</th>
<th>SRDA</th>
<th>PCA r=C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>0.725</td>
<td>0.770</td>
<td>0.735</td>
<td>0.708</td>
<td>0.698</td>
<td>0.707</td>
</tr>
<tr>
<td>COVTYPE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60%</td>
<td>0.720</td>
<td>0.755</td>
<td>0.719</td>
<td>0.704</td>
<td>0.685</td>
<td>0.683</td>
</tr>
<tr>
<td>30%</td>
<td>0.686</td>
<td>0.712</td>
<td>0.687</td>
<td>0.707</td>
<td>0.653</td>
<td>0.639</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>0.882</td>
<td>0.939</td>
<td>0.870</td>
<td>0.897</td>
<td>0.856</td>
<td>0.874</td>
</tr>
<tr>
<td>MNIST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60%</td>
<td>0.864</td>
<td>0.930</td>
<td>0.870</td>
<td>0.890</td>
<td>0.833</td>
<td>0.863</td>
</tr>
<tr>
<td>30%</td>
<td>0.848</td>
<td>0.916</td>
<td>0.850</td>
<td>0.881</td>
<td>0.786</td>
<td>0.842</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>0.946</td>
<td>0.946</td>
<td>0.950</td>
<td>0.699</td>
<td>0.936</td>
<td>0.940</td>
</tr>
<tr>
<td>LETTERS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60%</td>
<td>0.933</td>
<td>0.923</td>
<td>0.940</td>
<td>0.694</td>
<td>0.919</td>
<td>0.913</td>
</tr>
<tr>
<td>30%</td>
<td><strong>0.905</strong></td>
<td>0.885</td>
<td><strong>0.917</strong></td>
<td>0.680</td>
<td><strong>0.893</strong></td>
<td>0.872</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100%</td>
<td>0.936</td>
<td><strong>0.966</strong></td>
<td>0.925</td>
<td>0.943</td>
<td>0.921</td>
<td>0.930</td>
</tr>
<tr>
<td>USPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60%</td>
<td>0.927</td>
<td><strong>0.957</strong></td>
<td>0.917</td>
<td><strong>0.939</strong></td>
<td>0.906</td>
<td>0.921</td>
</tr>
<tr>
<td>30%</td>
<td>0.910</td>
<td><strong>0.942</strong></td>
<td>0.903</td>
<td><strong>0.926</strong></td>
<td>0.884</td>
<td>0.903</td>
</tr>
</tbody>
</table>
Avg. Classification Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Covtype</th>
<th>Mnist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labeled samples (in thousands)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training samples (in thousands)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For all training sizes only 30% of instances are annotated
Avg. Training time

<table>
<thead>
<tr>
<th>Labeled samples (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covtype</td>
</tr>
<tr>
<td>Mnist</td>
</tr>
</tbody>
</table>

Training samples (in thousands)

For all training sizes only 30% of instances are annotated
Conclusions

We presented a method whose main characteristics are:

- Modeling a semantic low-space representation
- Preserving the separability of the original classes
- Ability to exploit unlabeled instances for modeling the manifold structure of the data
- Online formulation