Modelling Relational Data using Bayesian Clustered Tensor Factorization

I. Sutskever, R. Salakhutdinov and J. B. Tenenbaum

NIPS 23, 2010

Presented by Esther Salazar
Duke University

August 26, 2011
Summary

- The authors consider the problem of learning probabilistic models for complex relational structures between various types of objects.

- Cluster based models → easily interpretable representations
  factorization-based approaches → better predictive performance on large data sets

- **Proposed model:** Bayesian Clustered Tensor Factorization (BCTF)
Summary

- The authors consider the problem of learning probabilistic models for complex relational structures between various types of objects.

- Cluster based models → easily interpretable representations
  factorization-based approaches → better predictive performance on large data sets

- **Proposed model:** Bayesian Clustered Tensor Factorization (BCTF)

**Relational data:**

\[(\text{object}, \text{relation}, \text{object})\]

**Examples:** (cup, used-for, drinking), (cup, can-contain, water), (cup, can-contain, coffee), (glass, can-contain, juice), (glass, can-contain, wine), and so on, we might also infer the propositions (glass, used-for, drinking) and (cup, can-contain, wine).
Suppose that we have a fixed set of objects \( O = \{a_1, \ldots, a_N\} \) and a fixed finite set of relations \( R = \{r_1, \ldots, r_M\} \).

Probabilistic models define a joint distribution over the truth values of all conceivable relations 
\[(a, r, b)\]

For binary variables: \( T(a, r, b) \in \{0, 1\} \) where \( a \) and \( b \) are objects, \( r \) is a relation and \( T(a, r, b) \) determines whether the relation \((a, r, b)\) is true.

Given a set of true relations \( S = \{(a, r, b)\} \) the model predicts that a new relation \((a, r, b)\) is true with probability
\[P(T(a, r, b) = 1 | S)\]
Bayesian Clustered Tensor Factorization (BCTF)

- For each object $a \in O$ the model maintains two vectors $a_L, a_R \in \mathbb{R}^d$
- For each relation $r \in R$ the model maintains a matrix $R \in \mathbb{R}^{d \times d}$
- Given $\theta = \{a_L, a_R, R\}$, the model independently chooses the truth-value of each relation $(a, r, b)$ from the distribution

$$P(T(a, r, b) = 1|\theta) = 1/(1 + \exp(-a_L^T R b_R))$$

Figure 1: A schematic diagram of the model, where the arcs represent the object clusters and the vectors within each cluster are similar. The model predicts $T(a, r, b)$ with $a_L^T R b_R$. 
Bayesian Clustered Tensor Factorization (BCTF)

Model
The observed data $\text{obs}$ is set of triples and their truth values $\{(a, r, b), t\}$. The model is defined as:

$$P(\text{obs}, \theta, c, \alpha, \alpha_{DP}) = P(\text{obs}|\theta, \sigma^2)P(\theta|c, \alpha)P(c|\alpha_{DP})P(\alpha_{DP}, \alpha, \sigma^2)$$

where $c = \{c_{obj}, c_{rel}\}$ contains the cluster assignments (partitions) of the objects and the relations; $\theta = \{a_L, a_R, R\}$ distributed representations of the objects and the relations, and $\{\sigma^2, \alpha, \alpha_{DP}\}$ the model parameters

$$P(\text{obs}|\theta) = \prod_{\{(a,r,b),t\}\in \text{obs}} \mathcal{N}(t|a_L^\top Rb_R, \sigma^2)$$

$$P(c|\alpha_{DP}) = \text{CRP}(c_{obj}|\alpha_{DP})\text{CRP}(c_{rel}|\alpha_{DP})$$
Bayesian Clustered Tensor Factorization (BCTF)

Defining $P(\theta|c, \alpha)$ takes a little more work. Given the partitions, the sets of parameters $\{a_L\}$, $\{a_R\}$, and $\{R\}$ become independent, so

$$P(\theta|c, \alpha) = P(\{a_L\}|c_{obj}, \alpha_{obj})P(\{a_R\}|c_{obj}, \alpha_{obj})P(\{R\}|c_{rel}, \alpha_{rel})$$

The distribution over the relation-vectors is given by

$$P(\{R\}|c_{rel}, \alpha_{rel}) = \prod_{k=1}^{|c_{rel}|} \int_{\mu, \Sigma} \prod_{i: c_{rel}, i = k} \mathcal{N}(R_i | \mu, \Sigma) \, dP(\mu, \Sigma | \alpha_{rel})$$

where $|c_{rel}|$ is the number of clusters in the partition $c_{rel}$. This is precisely a Dirichlet process mixture model [13]. We further place a Gaussian-Inverse-Gamma prior over $(\mu, \Sigma)$:

$$P(\mu, \Sigma | \alpha_{rel}) = P(\mu | \Sigma)P(\Sigma | \alpha_{rel}) = \mathcal{N}(\mu | 0, \Sigma) \prod_{d'} IG(\sigma_{d'}^2 | \alpha_{rel}, 1)$$

$$\propto \exp \left( - \sum_{d'} \frac{\mu_{d'}^2}{2} + 1 \right) \prod_{d'} (\sigma_{d'}^2)^{-0.5 - \alpha_{rel} - 1}$$

Finally, they place an improper $P(x) \propto x^{-1}$ scale-uniform prior over each hyperparameter independently.
Inference via MCMC

Before MCMC, find a MAP estimate then used it to initialize the Markov chain.

For each step of the Markov chain:

1. Given $\theta = \{a_L, a_R, R\}$, the chain updates $c = \{c_{obj}, c_{rel}\}$ using a collapsed Gibbs sampling sweep and a step of the split-and-merge algorithm.

2. Markov chain samples the parameter $\{a_L\}$ given $\{a_R\}$, $\{R\}$ and the cluster posterior means and covariances (tractable since the conditional dist. over the object vector $\{a_L\}$ is Gaussian).
Evaluation

The authors show that the BCTF model has excellent predictive power and that its finds interpretable clusters.

They compare the performance of the model to (i) the Infinite Relational Model (IRM), (ii) the Multiple Relational Clustering (MRC), (iii) its counterpart a Bayesian Tensor Factorization (BTF) (all the objects belong to a single cluster).

1. The animals dataset consists of 50 animals and 85 binary attributes. The dataset is a fully observed matrix—so there is only one relation.

2. The kinship dataset consists of kinship relationships among the members of the Alyawarra tribe [4]. The dataset contains 104 people and 26 relations. This dataset is dense and has 104⋅26=218216 observations, most of which are 0.

3. The UML dataset [11] consists of a 135 medical terms and 49 relations. The dataset is also fully observed and has 135⋅49=893025 (mostly 0) observations.

4. The MovieLens [15] dataset consists of 1000209 observed integer ratings of 6041 movies on a scale from 1 to 5, which are rated by 3953 users. The dataset is 95.8% sparse.

5. The Conceptnet dataset [10] is a collection of common-sense assertions collected from the web. It consists of about 112135 “common-sense” assertions such as (hockey, is-a, sport). There are 19 relations and 17571 objects. To make our experiments faster, we used only the 7000 most frequent objects, which resulted in 82062 true facts. For the negative data, we sampled twice as many random object-relation-object triples and used them as the false facts. As a result, there were 246186 binary observations in this dataset. The dataset is 99.9% sparse.
Comparisons

<table>
<thead>
<tr>
<th>algorithm</th>
<th>animals RMSE</th>
<th>animals AUC</th>
<th>kinship RMSE</th>
<th>kinship AUC</th>
<th>UML RMSE</th>
<th>UML AUC</th>
<th>movielens RMSE</th>
<th>movielens AUC</th>
<th>conceptnet RMSE</th>
<th>conceptnet AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP$_{20}$</td>
<td>0.467</td>
<td>0.78</td>
<td>0.122</td>
<td>0.82</td>
<td>0.033</td>
<td>0.96</td>
<td>0.899</td>
<td>–</td>
<td>0.536</td>
<td>0.57</td>
</tr>
<tr>
<td>MAP$_{40}$</td>
<td>0.528</td>
<td>0.68</td>
<td>0.110</td>
<td>0.90</td>
<td>0.024</td>
<td>0.98</td>
<td>0.933</td>
<td>–</td>
<td>0.614</td>
<td>0.48</td>
</tr>
<tr>
<td>BTF$_{20}$</td>
<td>0.337</td>
<td>0.85</td>
<td>0.122</td>
<td>0.82</td>
<td>0.033</td>
<td>0.96</td>
<td>0.835</td>
<td>–</td>
<td>0.275</td>
<td>0.93</td>
</tr>
<tr>
<td>BCTF$_{20}$</td>
<td><strong>0.331</strong></td>
<td><strong>0.86</strong></td>
<td>0.122</td>
<td>0.82</td>
<td>0.024</td>
<td>0.98</td>
<td><strong>0.834</strong></td>
<td>–</td>
<td>0.278</td>
<td>0.93</td>
</tr>
<tr>
<td>BTF$_{40}$</td>
<td>0.338</td>
<td><strong>0.86</strong></td>
<td><strong>0.108</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.024</strong></td>
<td><strong>0.98</strong></td>
<td>0.836</td>
<td>–</td>
<td><strong>0.260</strong></td>
<td><strong>0.94</strong></td>
</tr>
<tr>
<td>BCTF$_{40}$</td>
<td>0.336</td>
<td><strong>0.86</strong></td>
<td><strong>0.108</strong></td>
<td><strong>0.90</strong></td>
<td><strong>0.024</strong></td>
<td><strong>0.98</strong></td>
<td>0.836</td>
<td>–</td>
<td><strong>0.260</strong></td>
<td><strong>0.94</strong></td>
</tr>
<tr>
<td>IRM [8]</td>
<td>0.382</td>
<td>0.75</td>
<td>0.140</td>
<td>0.66</td>
<td>0.054</td>
<td>0.70</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MRC [9]</td>
<td>–</td>
<td>0.81</td>
<td>–</td>
<td>0.85</td>
<td>–</td>
<td><strong>0.98</strong></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 1: A quantitative evaluation of the algorithms using 20 and 40 dimensional vectors. We report the performance of the following algorithms: the MAP-based Tensor Factorization, the Bayesian Tensor Factorization (BTF) with MCMC (where all objects belong to a single cluster), the full Bayesian Clustered Tensor Factorization (BCTF), the IRM [8] and the MRC [9].
Results

Figure 2: Results on the Animals dataset. **Left:** The discovered clusters. **Middle:** The biclustering of the features. **Right:** The covariance of the distributed representations of the animals (bottom) and their attributes (top).
Results

![Image of results](image)

2 Amino Acid, Peptide, or Protein, Biomedical or Dental Material, Carbohydrate, ...
3 Amphibian, Animal, Archaean, Bird, Fish, Human, ...
4 Antibiotic, Biologically Active Substance, Enzyme, Hazardous or Poisonous Substance, Hormone, ...
5 Biologic Function, Cell Function, Genetic Function, Mental Process, ...
6 Classification, Drug Delivery Device, Intellectual Product, Manufactured Object, ...
7 Body Part, Organ, Cell, Cell Component, ...
8 Alga, Bacterium, Fungus, Plant, Rickettsia or Chlamydia, Virus
9 Age Group, Family Group, Group, Patient or Disabled Group, ...
10 Cell / Molecular Dysfunction, Disease or Syndrome, Model of Disease, Mental Dysfunction, ...
11 Daily or Recreational Activity, Educational Activity, Governmental Activity, ...
12 Environmental Effect of Humans, Human-caused Phenomenon or Process, ...
13 Acquired Abnormality, Anatomical Abnormality, Congenital Abnormality, Injury or Poisoning
14 Health Care Related Organization, Organization, Professional Society, ...

Figure 4: Results on the medical UML dataset. **Left:** The covariance of the distributed representations \( \{ a_L \} \) learned for each object. **Right:** The inferred clusters, along with the biclustering of a subset of the relations.
Results

Figure 5: Results on the Movielens dataset. **Left:** The covariance between the movie vectors. **Right:** The inferred clusters.
Discussion

- They introduced a new method for modelling relational data which is able to both discover meaningful structure and generalize well.

- A surprising aspect of the Bayesian model is the ease with which it worked after automatic hyperparameter selection was implemented.

- The model performs well even when the initial MAP estimate is very poor. This is particularly important for large sparse datasets, since finding a good MAP estimate requires careful cross-validation.

- Careful hyperparameter selection can be very labour-expensive because it requires careful training of a large number of models.