Block-LDA: Jointly modeling entity-annotated text and entity-entity links

Ramnath Balasubramanyan RBALASUB@CS.CMU.EDU
Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213 USA

William W. Cohen WCOHEN@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213 USA

Abstract

We present a model that improves entity entity link modeling in a mixed membership stochastic block model, by jointly modeling links with text about the entities that are linked in the relational data. The model also correspondingly improves the modeling of text annotated with entities using externally supplied entity-entity relations. We apply the model to a protein-protein interaction (PPI) dataset supplemented by a corpus of abstracts of scientific publications annotated with the proteins in the PPI dataset. Evaluation of the model using functional category prediction of proteins and perplexity shows improvements when joint modeling is used over baselines that uses only link or text information.

1. Introduction

The task of modeling relational information among entities is a commonly encountered problem. In social networks for instance, people list other people as friends in a social network and we might want to identify sub-communities from the social network. In the biological domain, proteins interact with other proteins and we would like to discover hidden attributes of proteins based on the observed pairwise interactions. Mixed membership stochastic block models (MMSB) (Airoldi et al., 2008; Parkkinen et al., 2009) approach the problem by assuming that nodes in the graph representing entities, belong to latent blocks with mixed membership, effectively capturing the notion that entities may arise from different sources and have different roles.

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Haifa, Israel, 2010.

In a parallel area of active research, models like Latent Dirichlet Allocation (Blei et al., 2003)(LDA) and others that extend it model text documents in a corpus as arising from a mixtures of latent topics. In such models, words in a document are potentially generated from different topics using topic specific distributions. Extensions to LDA proposed in (Erosheva et al., 2004; Griffiths & Steyvers, 2004) additionally model other metadata in documents such as authors and annotated entities by treating latent topics as sets of distributions, one each for every type of data in the documents. For instance, when modeling scientific publications from the biological domain, a latent topic could have a word distribution, author distribution and a protein mention distribution. We refer to this model as Link LDA following the convention established in (Nallapati et al., 2008).

In this paper, we present a model, **Block-LDA**, that jointly generates text documents annotated with entities and relational data containing links between pairs of entities. The model merges the idea of latent topics in topic models with blocks in stochastic block models. The joint modeling permits sharing of information about the latent topics between the network structure and text, resulting in more coherent topics. Co-occurrence patterns in entities and words related to them in the the text aid the modeling of links in the graph and the structure of entity relations provide clues about topics in the text, creating a symbiotic relationship. We also demonstrate a method to perform approximate inference in the model using a collapsed Gibbs sampler since exact inference in the joint model in intractable.

The Nubbi model (Chang et al., 2009) tackles a related problem where entity relations are discovered from textual data. The Topic-Link LDA model (Liu et al., 2009) deals with another related problem of modeling

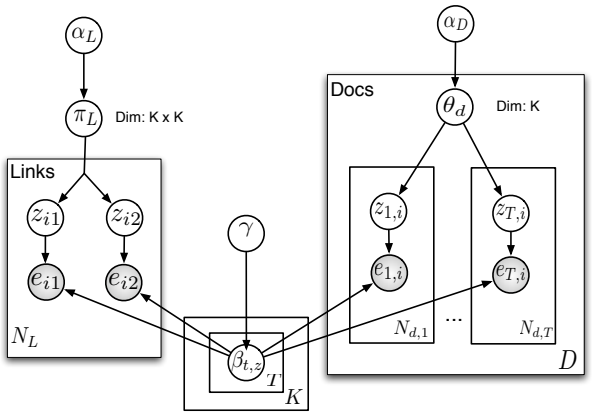

Figure 1. Block-LDA

author communities (or entity communities more generally), and text simultaneously. These models differ from the model presented in this paper in that they do not consider existing known relations between entities. Rather, they automatically discover relations in entities from the corpus. Additionally, the location of entities in text is key in the Nubbi model to build contexts in which entities involved in relations occur. In Block-LDA, the entities tagged in the document are separate from the text and are supplied as metadata. Pairwise-Link-LDA(Nallapati et al., 2008) also combines MMSB with LDA. Links in this model are however between entire documents and do not pertain to specific entities in the documents. A MMSB style treatment is given to links between documents. The Group-Topic model (Wang et al., 2006) addresses another related task of modeling events about entities and text about the event. The text in this model is however associated with events, which differs from the standalone documents mentioning entities considered by Block-LDA.

2. Block-LDA

The Block-LDA model (plate diagram in Fig 1) enables sharing of information between the component on the left that models links between pairs of entities represented as edges in a graph with a block structure, and the component on the right that models text documents, through shared latent topics. More specifically, the distribution over the entities of the type that are linked in the relational data is shared between the block model and the text model.

The component on the right, which is an extension of the Latent Dirichlet Allocation models documents as a set of "bag of entities", each bag corresponding to a particular type of entity. Every entity type has a topic wise multinomial distribution over the set of entities that can occur as an instance of the entity type.

The other component in the figure is a generative model for graphs representing relational data with an underlying block structure derived from the sparse block model introduced in (Parkkinen et al., 2009). Vertices in the graph representing entities have mixed memberships in topics, and edges(links) arise from a multinomial defined over the Cartesian product of topics. The linked entities are subsequently generated from topic specific entity distributions conditioned on the topic pair sampled for the edge. In contrast to MMSB, only observed links are sampled making this model suitable for sparse graphs.

Let K be the number of latent topics we wish to recover. Assuming documents consist of T different types of entities (i.e. each document contains T bags of entities), and that links in the graph are between entities of type t_l, the generative process is as follows.

First generate topics:

- For each type $t \in 1, \ldots, T$, and topic $z \in 1, \ldots, K$, sample $\beta_{t,z} \sim \text{Dirichlet}(\gamma)$.

Then generate documents. For every document d:

- Sample $\theta_d \sim \text{Dirichlet}(\alpha_D)$ where θ_d is the topic mixing distribution for the document.
- For each type t and its associated set of entity mentions $e_{t,i}$, $i \in \{1, \cdots, N_{d,t}\}$:
  - Sample a topic $z_{t,i} \sim \text{Multinomial}(\theta_d)$
  - Sample an entity $e_{t,i} \sim \text{Multinomial}(\beta_{t,z_{t,i}})$

Finally generate the link matrix of entities of type t_l:

- Sample $\pi_L \sim \text{Dirichlet}(\alpha_L)$ where π_L describes the distribution over the Cartesian product of topics for links in the dataset.
- For every link $e_{i1} \rightarrow e_{i2}$:
  - Sample a topic pair $\langle z_{i1}, z_{i2} \rangle \sim \text{Multinomial}(\pi_L)$
  - Sample $e_{i1} \sim \text{Multinomial}(\beta_{t_l,z_{i1}})$
  - Sample $e_{i2} \sim \text{Multinomial}(\beta_{t_l,z_{i2}})$

Due to the intractability of exact inference in the Block-LDA model, a collapsed Gibbs sampler is used to perform approximate inference. It samples a latent topic for an entity mention of type t in the text corpus conditioned on the assignments to all other entity mentions using the following expression (after collapsing θ_D):

$$p(z_{t,i} = z | e_{t,i}, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{e}^{\neg \mathbf{i}}, \alpha_D, \gamma)$$
$$\propto (n_{dz}^{\neg i} + \alpha_D) \frac{n_{zte_{t,i}}^{\neg i} + \gamma}{\sum_{e'} n_{zte'}^{\neg i} + |E_t|\gamma}$$

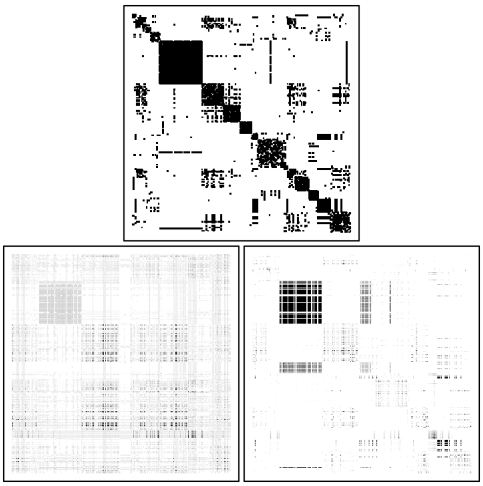

Figure 2. (Top) MIPS protein protein interactions. (Bottom left) Inferred using the Sparse block model. (Bottom right) Inferred using Block-LDA.

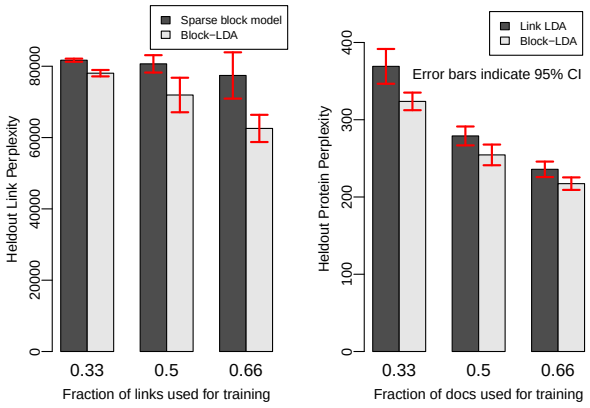

Figure 3. Gain in perplexity through joint modeling

Similarly, we sample a topic pair for every link conditional on topic pair assignments to all other links after collapsing π_L using the expression:

$$p(\mathbf{z_i} = \langle z_1, z_2\rangle | \langle e_{i1}, e_{i2}\rangle, \mathbf{z}^{\neg \mathbf{i}}, \langle \mathbf{e_1}, \mathbf{e_2}\rangle^{\neg \mathbf{i}}, \alpha_L, \gamma)$$
$$\propto \left(n^{L\neg i}_{\langle z_1,z_2\rangle} + \alpha_L\right) \times$$
$$\frac{(n^{\neg i}_{z_1 t_l e_{i1}}+\gamma)(n^{\neg i}_{z_2 t_l e_{i2}}+\gamma)}{\left(\sum_e n^{\neg i}_{z_1 t_l e}+|E_{t_l}|\gamma\right)\left(\sum_e n^{\neg i}_{z_2 t_l e}+|E_{t_l}|\gamma+\delta_{z_1,z_2}\right)}$$

where δ_{z_1,z_2} is 1 when $z_1 = z_2$ and 0 otherwise. E_t refers to the set of all entities of type t. The n's are counts of observations in the training set.

3. Dataset

The Munich Institute for Protein Sequencing (MIPS) database includes a hand-crafted collection of protein interactions covering 8000 protein complex associations in yeast. We use a subset of this collection containing 844 proteins, for which all interactions were hand-curated (See top panel of Fig 2). The MIPS institute also provides a set of functional annotations for each protein which are organized in a tree, with 15 nodes high-level functions at the first level. The 844 proteins participating in interactions are mapped to these 15 functional categories with an average of 2.5 annotations per protein. (more details about the dataset are in (Airoldi et al., 2008))

In addition to the MIPS PPI data, we use a text corpus that is derived from the repository of scientific publications at PubMed Central. PubMed is a free, open-access on-line archive of over 18 million biological abstracts and bibliographies, including citation lists, for papers published since 1948 (U.S. National Library of Medicine 2008). The subset we work with consists of approximately 40,000 publications about the yeast organism that have been curated in the Saccharomyces Genome Database (SGD)[1], with various types of information concerning the organism *Saccharomyces cerevisiae*, including tags of genes and proteins which the publications discuss. We select 15,776 such documents tagged with proteins from the MIPS database. The publications in this set were written by a total of 47,215 authors. We tokenize the titles and abstracts based on white space, lowercase all tokens and eliminate stopwords. Low frequency (< 5 occurrences) terms are also eliminated. The vocabulary contains 45,648 words.

4. Results

We perform two sets of experiments with the PPI+SGD dataset. The SGD text data has 3 types of entities in each document - words, authors and protein annotations with the PPI data linking proteins. In the first set of experiments, we evaluate the model using perplexity of heldout protein-protein interactions using increasing amounts of the PPI data for training. All the 15,773 documents in the SGD dataset are used when textual information is used. When text is not used, the model is equivalent to using only the left half of Fig 1. The bottom left and right panels in Fig 2 shows the probability of protein protein interactions recovered using the sparse block model and using Block-LDA respectively. In the other set of experiments, we evaluate the model using protein perplexity in heldout text using progressively increasing amounts of text as training data. All the links in the PPI dataset are used in these experiments when link data is used. When link data is not used, the model

[1] http://www.yeastgenome.org

reduces to Link LDA. In all experiments, the Gibbs sampler is run until the held out perplexity stabilizes to a nearly constant value ($\approx$ 80 iterations)

Fig 3 shows the gains in perplexity in the two sets of experiments with different amounts of training data. The perplexity values are averaged over 10 runs. In both sets of experiments, it can be seen that Block-LDA results in lower perplexities than using links/text alone. These results indicate that co-occurrence patterns of proteins in text contain information about protein interactions which Block-LDA is able to utilize through joint modeling. Our conjecture is that the protein co-occurrence information in text is a noisy approximation of the PPI data.

4.1. Functional category prediction

Proteins are identified are identified as belonging to multiple functional categories in the MIPS dataset, as described in Section 3. We use Block-LDA and baseline methods to predict proteins' functional categories and evaluate it by comparing it to the ground truth in the MIPS dataset. A model is first trained with K set to 15 topics to hopefully recover the 15 top level functional categories of proteins. Every topic that is returned consists of a set of multinomials including $\beta_{\mathbf{t}_1}$, the topic wise distribution over all proteins. The values of $\beta_{\mathbf{t}_1}$ are thresholded such that the top $\approx 16\%$ (the density of the protein-function matrix) of entries are considered as a positive prediction that the protein falls in the functional category corresponding to the latent topic. To determine the mapping of latent topic to functional category, 10% of the proteins are used in a procedure that greedily finds the alignment resulting in the best accuracy, as described in (Airoldi et al., 2008). The precision, recall and F_1 scores of the different models in predicting the right functional categories for proteins are shown in Table 4.1. For the random baseline, every protein-functional category pair is randomly deemed to be 0 or 1 with the Bernoulli probability of an association being proportional to the ratio of 1's observed in the protein-functional category matrix in the MIPS dataset. In the MMSB approach, induced latent blocks are aligned to functional categories as described in (Airoldi et al., 2008).

We see that the F_1 scores for the baseline sparse block model and MMSB are nearly the same and that combining text and links provides a significant boost to the F_1 score which suggests that protein co-occurrence patterns in the abstracts contains information about functional categories as is also evidenced by the fairly good F_1 score obtained using Link LDA which uses only documents. All the methods considered outperform the random baseline.

Method	F_1	Precision	Recall
Block-LDA	**0.249**	0.247	0.250
Sparse Block model	0.161	0.224	0.126
Link LDA	0.152	0.150	0.155
MMSB	0.165	0.166	0.164
Random	0.145	0.155	0.137

Table 1. Functional category prediction

5. Conclusion

We proposed a model that jointly models links between entities and text annotated with entities that permits co-occurrence information in text to influence link modeling and vice versa. Our experiments show that joint modeling outperforms approaches that use only links/text when evaluated internally using perplexity and externally using protein functional category prediction.

References

Airoldi, Edoardo M., Blei, David M., Fienberg, Stephen E., and Xing, Eric P. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981—2014, September 2008.

Blei, D. M, Ng, A. Y, and Jordan, M. I. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Chang, J., Boyd-Graber, J., and Blei, D. M. Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 169178, 2009.

Erosheva, E., Fienberg, S., and Lafferty, J. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (Suppl 1):5220, 2004.

Griffiths, T. L. and Steyvers, M. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004. ISSN 0027-8424.

Liu, Yan, Niculescu-mizil, Alexandru, and Gryc, Wojciech. Topic-Link LDA: joint models of topic and author community. In Bottou, Lon and Littman, Michael (eds.), *Proceedings of the 26th International Conference on Machine Learning*, pp. 665672, Montreal, June 2009. Omnipress.

Nallapati, Ramesh M., Ahmed, Amr, Xing, Eric P., and Cohen, William W. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 542–550, Las Vegas, Nevada, USA, 2008. ACM.

Parkkinen, Juuso, Sinkkonen, Janne, Gyenge, Adam, and Kaski, Samuel. A block model suitable for sparse graphs. In *Proceedings of the 7th International Workshop on Mining and Learning with Graphs (MLG 2009)*, Leuven, 2009. Poster.

Wang, Xuerui, Mohanty, Natasha, and Mccallum, Andrew. Group and topic discovery from relations and their attributes. *In Advances in Neural Information Processing Systems 18*, pp. 1449—1456, 2006.