Group Sparse Topical Coding: From Code to Topic

Lu Bai, Jiafeng Guo, Yanyan Lan, Xueqi Cheng
Institute of Computing Technology, Chinese Academy of Sciences
{bailu, guojiafeng}@software.ict.ac.cn, {lanyanyan, cxq}@ict.ac.cn

1. MOTIVATION

PTM (Probabilistic Topic Model)
- Document is modeled as a meaningful low dimensional point in the topic simplex.
- Lack a mechanism to actively control the posterior sparsity of the desired representations.
- Does not allow the automatic detection of the number of topics.

NPM (Non-Probabilistic Model)
- Time to achieve sparsity by using some constraint like basis or other composite regularizer.
- Does not allow the automatic detection of the number of topics.

Group Sparse Topical Coding (GSTC)
- Produce a meaningful representation of a document.
- Control the sparsity of representation directly.
- Model learning efficiently.

2. BASIC IDEA

The meaning of document is composed of the meanings of words.
Restricting the topics of words in the document can impact the meanings of document implicitly.

3. MODEL DETAIL

3.1 Object function

\[ \min_{\beta} L(\beta) = -\log p(D|\beta) \]

3.2 Generating process of a document
- For each topic \( k \in \{1, \ldots, K\} \), sample a word code vector \( x_i \in \mathbb{R}^K \) from \( \text{Multinomial}(\beta_k) \).
- For each observed word \( n \in I \), for each topic \( k \in \{1, \ldots, K\} \), sample a latent word count \( w_{kn} \sim \text{Poisson}(\epsilon_k \beta_{kn}) \).
- Obtain the word count \( w_n = \sum_k w_{kn} \).

3.3 Parameter estimation
1. Fix \( \beta \), learn \( \epsilon_k \) for each document \( d \) (block coordinate descent are used for group sparsity).
2. Fix \( \epsilon_k \), learn \( \beta \) with projected gradient method.
3. Go to step 1 until convergence.

4. EXPERIMENTAL RESULTS

4.1 Dataset
- 20-newsgroup
  - 18,846 documents,
  - 20,214 distinct words
  - 20 related categories

4.2 Baseline methods
- LDA, NMF, STC

4.3 Evaluation
- Topic sparsity
- Train time
- Accuracy of document classification

5. CONCLUSIONS

Conclusion
- GSTC provides an elegant way to model topics concerning both sparsity and semantic representation.
- Experiments show the good performance of GSTC in meaningful compact latent representations and document classification.

Future work
- Consider the sparsity of dictionary.
- Develop a paralleled algorithm for large-scale applications.
- Extend the GSTC by integrating the discriminative features of document.