输入先验主题模型

Text Modeling using Unsupervised Topic Models and Concept Hierarchies, 2008

1. 问题

将非监督学习得到的主题(topic)与人类的先验概念(concept)结合.

主题是单词的多项分布; 概念是单词的集合.

2. 方法

基于LDA.

概念-主题模型(concept-topic model): 将C个概念转换成主题, 加入到T个模型主题中, 形成$T+C$个主题. 算法为

1. For each topic $t \in \{1, ..., T\}$, select a word distribution $\phi_t \sim \text{Dir}(\beta_\phi)$
2. For each concept $c \in \{1, ..., C\}$, select a word distribution $\psi_c \sim \text{Dir}(\beta_\psi)$ [1]
3. For each document $d \in \{1, ..., D\}$
   (a) Select a distribution over topics and concepts $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each word w of document d
      i. Select a component $z \sim \text{Mult}(\theta_d)$
      ii. If $z \le T$ generate a word from topic z, $w \sim \text{Mult}(\phi_z)$; otherwise generate a word from concept $c = z - T$, $w \sim \text{Mult}(\psi_c)$

如果$C=0$则退化成LDA.

分层概念-主题模型(hierachical concept-topic model): 概念通常存储为树结构, 而前面的概念-主题模型则忽视了这个结构. 引入一个开关分布$p(x|d)$, 决定单词从主题路径还是概念路径生成. 即$x=0$时, 由主题模型生成单词; $x=1$时, 从概念树上生成单词. LDA中条件概率变为

$$
\begin{aligned}
p(w|d) = P(x=0|d) \sum_t p(w|t)p(t|d) \\
+ P(x=1|d) \sum_c p(w|c)p(c|d)
\end{aligned}
$$

$$
\text{where} \quad p(c|d) = p(exit|c)p(c|parent(c))...p(.|root)
$$

生成算法变为

1. For each topic $t \in \{1, ..., T\}$, select a word distribution $\phi_t \sim \text{Dir}(\beta_\phi)$
2. For each concept $c \in \{1, ..., C\}$, select a word distribution $\psi_c \sim \text{Dir}(\beta_\psi)$ [2]
3. For each document $d \in \{1, ..., D\}$
   (a) Select a switch distribution $\xi_d \sim \text{Beta}(\gamma)$
   (b) Select a distribution over topics $\theta_d \sim \text{Dir}(\alpha)$
   (c) For each concept $c \in \{1, ..., C\}$
      i. Select a distribution over children of c, $\zeta_{cd} \sim \text{Dir}(\tau_c)$
   (d) For each word w of document d
      i. Select a binary switch variable $x \sim \text{Bernoulli}(\xi_d)$
      ii. If $x = 0$
         A. Select a topic $z \sim \text{Mult}(\theta_d)$
         B. Generate a word from topic z, $w \sim \text{Mult}(\phi_z)$
      iii. Otherwise, create a path starting at the root concept node, $\lambda_1 = 1$
         A. Repeat
            Select a child of node λ_j, $\lambda_{j+1} \sim \text{Mult}(\zeta_{\lambda_j d})$
            Until λ_{j+1} is an exit node
         B. Generate a word from concept $c = \lambda_j$, $w \sim \text{Mult}(\psi_c)$; set z to $T + c$

3. 实验

- 数据集

  文本数据来自TASA(Touchstone Applied Science Associates), 是教育相关数据, 包含37651个文档. 文章选取其中SCIENCE和SOCIAL STUDIES两个目录, 分别包含5356和10501个文档, 都有上百万个单词.

  概念数据来自: 1. CALD(Cambridge Advanced Learner's Dictionary), 包含2183个分层组织的语义目录, 从EVERYTHING概念下分的第二层由17个概念, 包含SCIENCE, SOCIETY等等, 一共下分到7层, 相应概念最少有54个单词, 最多有3074个单词. 2. ODP(Open Directory Project), 分层组织, SCIENCE子树下包含10817个概念.

- 结果

  比较各个模型的困惑度(perplexity).

  SCIENCE概念:

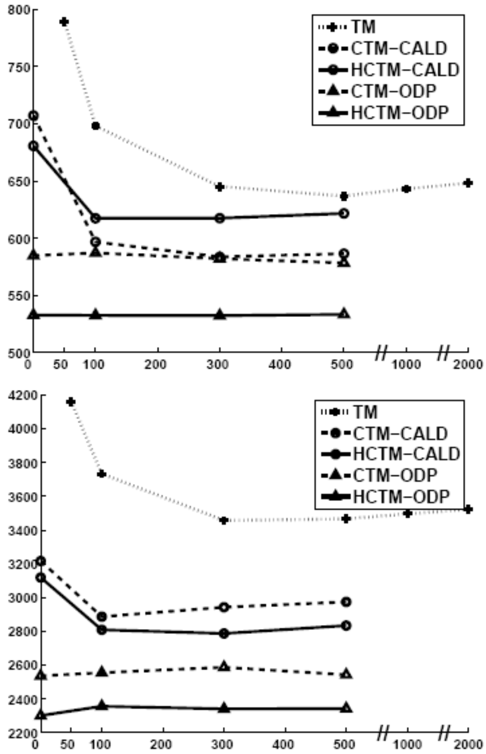


Figure 4: Comparing perplexity for TM, CTM and HCTM using training documents from science and testing on science (top) and social studies (bottom) as a function of number of topics

其中TM指LDA模型, CTM指概念-主题模型, HCTM指分层概念-主题模型.

SOCIAL STUDIES概念:

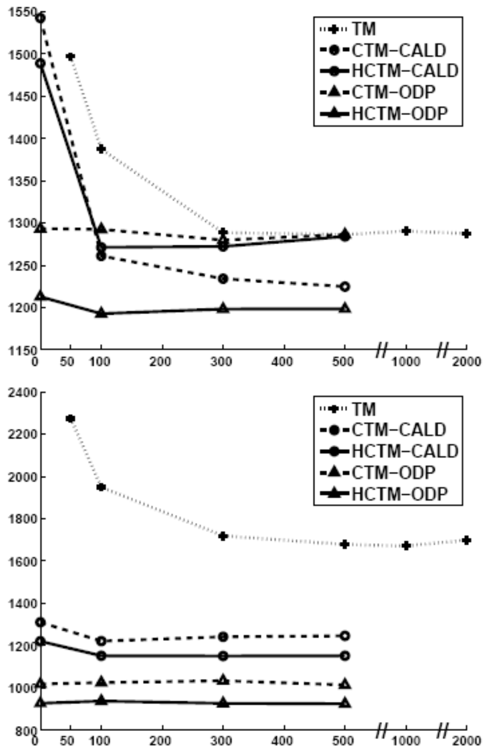


Figure 5: Comparing perplexity for TM, CTM and HCTM using training documents from social studies and testing on social studies (top) and science (bottom) as a function of number of topics

Seed-Guided Topic Model for Document Filtering and Classification, 2018

1. 问题

过滤不相关文档, 将相关文档分类.

2. 方法

基于LDA.

给出一组未标记文档, 每个类别设置相关的种子词, 根据文档主题与种子词的关系, 进行过滤与分类. 创新点是假设文档中主题是由一个"类别主题"(category-topic)和若干个"一般主题"(general-topic)组成, 利用少量的类别种子词自动与文档中某个"类别主题"(category-topic)建立联系, 可以代替海量的人工标记过程.

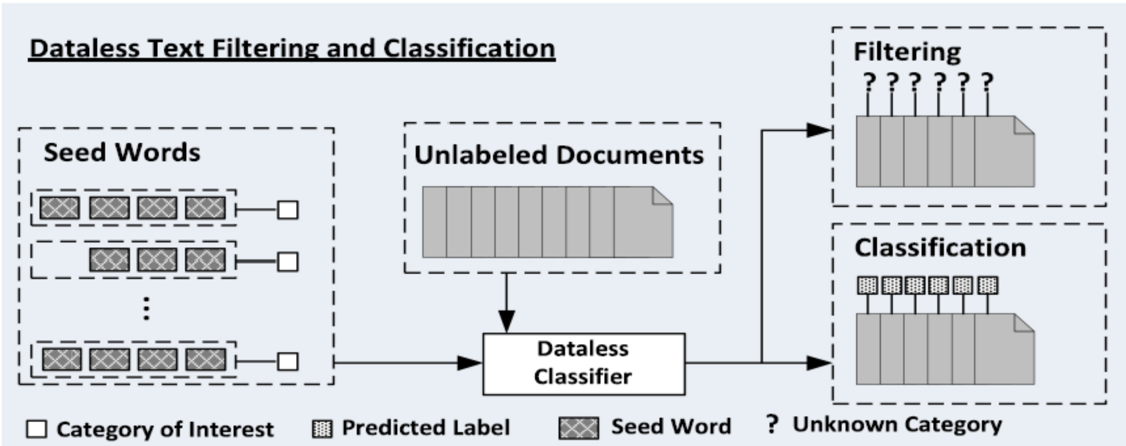


Fig. 2. Illustration for dataless filtering and classification.

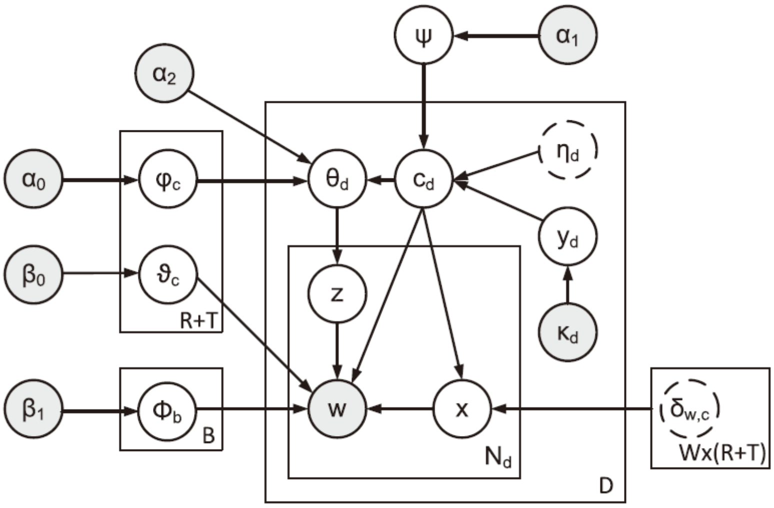


Fig. 4. Graphical representation of DFC. Since the variables η_d and $\delta_{w,c}$ are estimated based on the (pseudo) seed words of DFC, we therefore plot these two sets of variables in dotted circle.

3.实验

- 数据集

20-Newsgroup(20NG), 包含18846个文档, 可分为20个不同的类别.

Reuters-10, 原本的Reuters-21578包含21578个文档, 可分为135个类别. 本文使用其中规模最大的10个类别, 并且去除属于多个类别的文档, 剩下7285个文档.

- 结果

  种子词的取法分两种, 一种是直接从标签中得到的$\mathbb{S}^L$, 另一种是从描述中得到的$\mathbb{S}^D$. 比较各类方法的macro-F1指标.

  过滤与分类:

Table 4. Macro-F_1 of the Six Methods for Classification with Filtering, Where the Seed Words in $\mathbb{S}^L$ Are Used

Dataset	Classification task	DFC		TLC++	SSVM	MedLDA	sLDA	SVM
		Doc-Rel	Topic-Rel					
20NG	med	0.580†	**0.692**	0.004†	0.100	0.256†▼	0.701▲	0.190▼
	space	0.329†	**0.756**	0.027†	0.166	0.304†▼	0.763▲	0.255▼
	sci	**0.716**	0.291†	0.007†	0.298	0.426†▼	0.755†▲	0.370▼
	religion	**0.870**	0.432†	0.500†	0.391	0.390†▼	0.857▼	0.409▼
	med-space	0.595†	**0.791**	0.047†	0.133	0.332†▼	0.737†▼	0.222▼
	pc-mac	0.179†	**0.242**	0.019†	0.131	0.350†▲	0.362†▲	0.208▼
	politics-religion	**0.849**	0.506†	0.311†	0.359	0.564†▼	0.861▲	0.423▼
	politics-sci	0.602†	**0.684**	0.285†	0.313	0.581†▼	0.797†▲	0.404▼
	comp-religion-sci	**0.779**	0.585†	0.448†	0.378	0.637†▼	0.850†▲	0.418▼
	politics-rec-religion-sci	**0.839**	0.658†	0.578†	0.341	0.705†▼	0.848▲	0.407▼
	autos-motorcycles-baseball-hockey	0.839	**0.844**	0.262†	0.148	0.487†▼	0.782†▼	0.262▼
Reuters-10	gold	**0.848**	0.682†	0.000†	0.294	0.000†▼	0.620†▼	0.333▼
	earn	0.947	0.887†	0.943	0.419	0.489†▼	0.984†▲	0.822▼
	acq-earn	0.892†	0.274†	**0.940**	0.443	0.653†▼	0.980†▲	0.636▼
	coffee-gold-sugar	**0.915**	0.841†	0.093†	0.563	0.121†▼	0.764†▼	0.574▼
Avg		0.719	0.611	0.298	0.298	0.420	0.777	0.396

The best and second best results by dataless classifiers are highlighted in boldface and underlined, respectively, on each task. † indicates that the difference to the best dataless classifier is statistically significant at 0.05 level. ▲ and ▼ indicate that the supervised classifiers perform better or worse than the best dataless classifier, respectively. Avg: the averaged Macro-F_1 over all tasks.

Table 5. Macro-F_1 of the Six Methods for Classification with Filtering, Where the Seed Words in $\mathbb{S}^D$ Are Used

Dataset	Classification task	DFC		TLC++	SSVM	MedLDA	sLDA	SVM
		Doc-Rel	Topic-Rel					
20NG	med	**0.863**	0.820†	0.004†	0.159	0.256†▼	0.701†▼	0.190▼
	space	**0.861**	0.758†	0.027†	0.181	0.304†▼	0.763†▼	0.255▼
	sci	**0.716**	0.639†	0.007†	0.349	0.426†▼	0.755†▲	0.370▼
	religion	**0.872**	0.571†	0.500†	0.375	0.390†▼	0.857▼	0.409▼
	med-space	**0.866**	0.823†	0.047†	0.170	0.333†▼	0.737†▼	0.222▼
	pc-mac	**0.426**	0.283†	0.019†	0.174	0.350†▼	0.362†▼	0.208▼
	politics-religion	**0.899**	0.771†	0.311†	0.395	0.564†▼	0.861†▼	0.423▼
	politics-sci	**0.811**	0.693†	0.285†	0.382	0.581†▼	0.797▼	0.404▼
	comp-religion-sci	**0.793**	0.743†	0.448†	0.392	0.637†▼	0.850†▲	0.418▼
	politics-rec-religion-sci	**0.853**	0.777†	0.578†	0.387	0.705†▼	0.848▼	0.407▼
	autos-motorcycles-baseball-hockey	**0.864**	0.850	0.262†	0.266	0.487†▼	0.782†▼	0.262▼
Reuters-10	gold	**0.818**	0.620†	0.000†	0.167	0.000†▼	0.620†▼	0.333▼
	earn	0.908†	0.942†	**0.943**	0.775	0.489†▼	0.984†▲	0.822▼
	acq-earn	0.803†	0.919†	**0.940**	0.636	0.653†▼	0.980†▲	0.636▼
	coffee-gold-sugar	0.731†	**0.820**	0.093†	0.310	0.121†▼	0.764†▼	0.574▼
Avg		0.806	0.735	0.298	0.341	0.420	0.777	0.396

The best and second best results are highlighted in boldface and underlined, respectively, on each task. † indicates that the difference to the best dataless classifier is statistically significant at 0.05 level. ▲ and ▼ indicate that the supervised classifiers perform better or worse than the best dataless classifier, respectively. Avg: the averaged Macro-F_1 over all tasks.

只分类不过滤:

Table 6. Macro-F_1 of the Eight Methods for Classification without Filtering, Where the Seed Words in $\mathbb{S}^L$ Are Used

Dataset	Classification task	DFC		Ge-FL	DescLDA	SNB-EM	TLC++	MedLDA	sLDA	SVM
		Doc-Rel	Topic-Rel							
20NG	med-space	0.967	**0.972**	0.712†	0.877†	0.897	0.938†	0.975▲	0.910†▼	0.976▲
	pc-mac	**0.902**	0.678†	0.491†	0.688†	0.895	0.685†	0.881†▼	0.735†▼	0.925▲
	politics-religion	0.907	0.506†	0.684†	0.888†	0.894	**0.911**	0.949†▲	0.925†▲	0.954▲
	politics-sci	**0.960**	0.746†	0.750†	0.624†	0.846	0.906†	0.941†▼	0.930†▼	0.971▲
	comp-religion-sci	**0.918**	0.857†	0.709†	0.559†	0.907	0.817†	0.930▲	0.900▼	0.936▲
	politics-rec-religion-sci	**0.919†**	0.742†	0.719†	0.514†	0.768	0.834†	0.932▲	0.823†▼	0.941▲
	autos-motorcycles-baseball-hockey	0.936†	**0.957**	0.849†	0.531†	0.715	0.734†	0.962▲	0.894†▼	0.957
	All 20 categories	**0.662†**	0.572†	0.320†	0.632†	0.461	0.510†	0.705†▲	0.633†▼	0.820▲
Reuters-10	All 10 categories	**0.701†**	0.496†	0.667†	0.317†	0.529	0.506†	0.562†▼	0.754†▲	0.932▲
Avg		0.875	0.725	0.656	0.626	0.768	0.760	0.871	0.834	0.935

The best and second best results by dataless classifiers are highlighted in boldface and underlined, respectively, on each task. † indicates that the difference to the best dataless classifier is statistically significant at 0.05 level. ▲ and ▼ indicate that the supervised classifiers perform better or worse than the best dataless classifier, respectively. Avg: the averaged Macro-F_1 over all tasks.

Table 7. Macro-F_1 of the Eight Methods for Classification without Filtering, Where the Seed Words in $\mathbb{S}^D$ Are Used

Dataset	Classification task	DFC		Ge-FL	DescLDA	SNB-EM	TLC ++	MedLDA	sLDA	SVM
		Doc-Rel	Topic-Rel							
20NG	med-space	0.972	**0.979**	0.935†	0.977	0.967	0.938†	0.975▲	0.910†▼	0.976▼
	pc-mac	**0.936**	0.416†	0.705†	0.694†	0.876	0.685†	0.881†▼	0.735†▼	0.925▼
	politics-religion	**0.952**	0.935	0.883†	0.900†	0.939	0.911†	0.949▼	0.925†▼	0.954▲
	politics-sci	**0.962**	0.912†	0.889†	0.912†	0.941	0.906†	0.941†▼	0.930†▼	0.971▲
	comp-religion-sci	**0.923**	0.861†	0.828†	0.498†	0.919	0.817†	0.930▲	0.900†▼	0.936▲
	politics-rec-religion-sci	**0.941**	0.914†	0.827†	0.782†	0.917	0.834†	0.932▼	0.823†▼	0.941
	autos-motorcycles-baseball-hockey	**0.977**	0.957	0.673†	0.713†	0.938	0.734†	0.962▼	0.894†▼	0.957▼
	All 20 categories	**0.739**	0.728	0.590†	0.663†	0.678	0.510†	0.705†▼	0.633†▼	0.820▲
Reuters-10	All 10 categories	**0.822**	0.791†	0.776†	0.800†	0.778	0.506†	0.562†▼	0.754†▼	0.932▲
Avg		0.914	0.833	0.790	0.771	0.884	0.760	0.871	0.834	0.935

The best and second best results by dataless classifiers are highlighted in boldface and underlined, respectively, on each task. † indicates that the difference to the best dataless classifier is statistically significant at 0.05 level. ▲ and ▼ indicate that the supervised classifiers perform better or worse than the best dataless classifier, respectively. Avg: the averaged Macro-F_1 over all tasks.

- Source-LDA: Enhancing probabilistic topic models using prior knowledge sources, 2017

1. 问题

文档分类. 一般有两种方法: 1. 后处理过程, 即确认了文档主题分布之后, 选择先验知识库(prior knowledge base)中最接近的主题进行标记; 2. 监督主题模型, 将主题限制到预先定义好的集合中, 即这个集合的每个主题的单词分布预先给定. 前者的缺点是标记的主题不能准确表示文档的单词分布, 后者难以发现未知主题. 文章给出了一个平衡的半监督学习方法.

2. 方法

基于LDA.

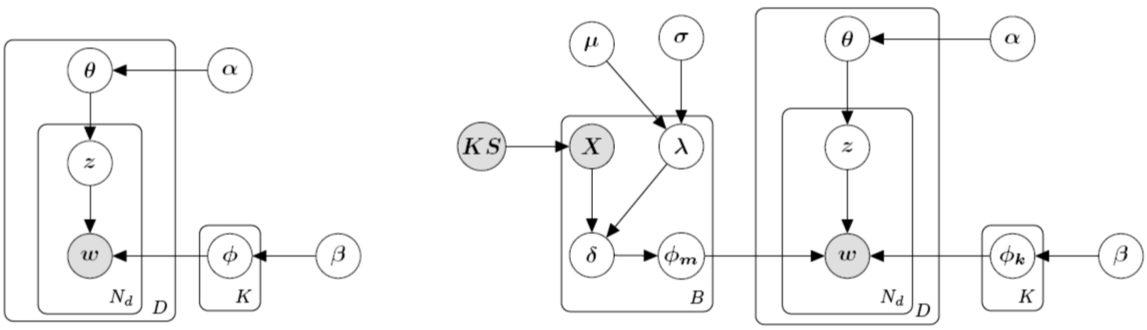
LDA算法:

1. For each of the K topics ϕ_k:
2.   Choose $\phi_k \sim \mathrm{Dir}(\beta)$
3. For each of the D documents d:
4.   Choose $N_d \sim \mathrm{Poisson}(\xi)$
5.   Choose $\theta_d \sim \mathrm{Dir}(\alpha)$
6.   For each of the N_d words $w_{n,d}$:
7.   Choose $z_{n,d} \sim \mathrm{Multinomial}(\theta)$
8.   Choose $w_{n,d} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$

Source-LDA算法:

1. For each of the T topics ϕ_t:
2.   **if** $t \leq K$ **then**
3.   Choose $\phi_t \sim \mathrm{Dir}(\beta)$
4.   **else**
5.   Choose $\lambda_t \sim \mathcal{N}(\mu, \sigma)$
6. $\delta_t \leftarrow [(X_{t,1})^{g(\lambda_t)}, (X_{t,2})^{g(\lambda_t)}, \ldots, (X_{t,V})^{g(\lambda_t)}]$
7.   Choose $\phi_t \sim \mathrm{Dir}(\delta_t)$
8. For each of the D documents d:
9.   Choose $N_d \sim \mathrm{Poisson}(\xi)$
10.   Choose $\theta_d \sim \mathrm{Dir}(\alpha)$
11.   For each of the N_d words $w_{n,d}$:
12.   Choose $z_{n,d} \sim \mathrm{Multinomial}(\theta)$
13.   Choose $w_{n,d} \sim \mathrm{Multinomial}(\phi_{z_{n,d}})$

其中第6,7行是为了引入主题分布的影响, 第2, 4行的条件语句是将已知主题和未知主题区别对待, 第5行引入的参数λ(需要调试的超参数)是为了增加松弛, 即允许一些从资源分布(source distribution)中带来的偏差. 可以看到, Source-LDA的改进只存在于主题分布ϕ_k的生成上.



(a) LDA   (b) Source-LDA

Fig. 1: Plate notation for LDA (a), and the proposed Source-LDA (b).

3. 实验

代码可见https://github.com/ucla-scai/Source-LDA.

- 数据集1

  手动生成的图:

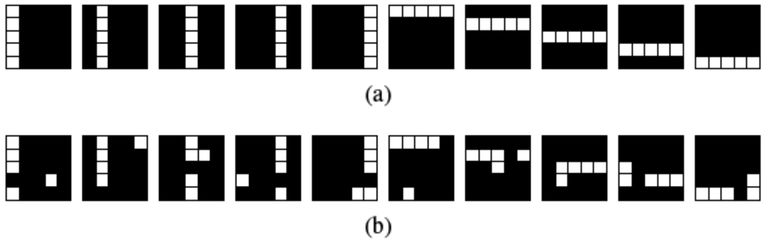


Fig. 5: A graphical representation of topics containing 1 word for the cell locations of row and column vectors in a 5 x 5 picture (a) and their augmented topics after swapping a random assigned word (pixel) with a random topic's assigned word (b).

得到的结果:

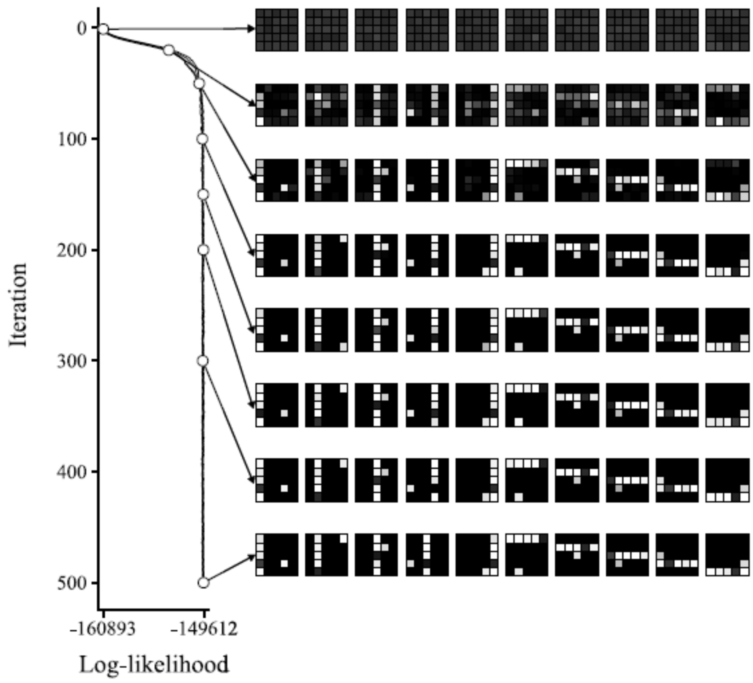


Fig. 6: Results from running Source-LDA for a corpus generated from topics in Figure 5(b) using a knowledge source of topics corresponding to Figure 5(a). Four separate runs are plotted to show the similarity of the log-likelihood relation to the iteration between the runs. The topics are shown visually at iteration 1, 20, 50, 100, 150, 200, 300 and 500 for a single run.

不同λ选取对精度的影响:

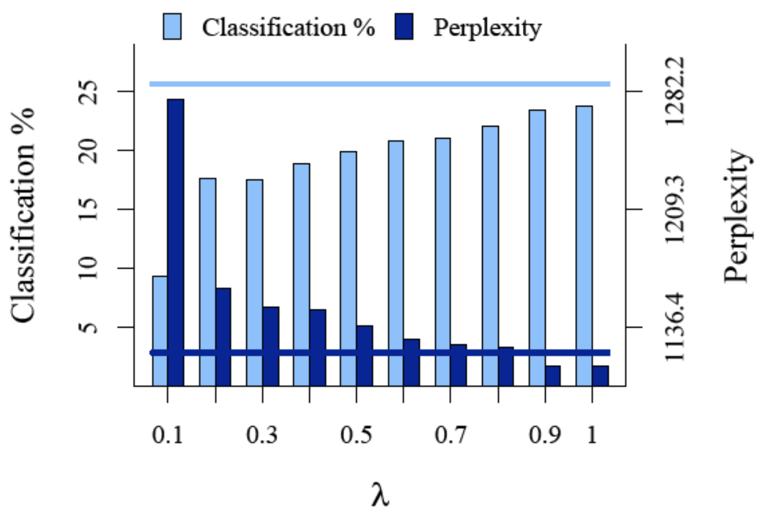


Fig. 7: Classification accuracy and perplexity values for fixed values of λ compared against the baseline values generated from a dynamic λ with a normal prior. The baseline values shown as lines represent the classification percentage of 25.7 and perplexity value of 1119.9

- 数据集2

  Reuters-21578.

  不同方法的比较:

Inventories			Natural Gas			Balance of Payments		
SRC-LDA	IR-LDA	CTM	SRC-LDA	IR-LDA	CTM	SRC-LDA	IR-LDA	CTM
inventory	systems	sales	gas	corp	gas	account	said	said
cost	products	year	natural	contract	said	surplus	public	june
stock	said	sold	used	company	total	deficit	state	april
accounting	information	retail	water	services	value	current	private	beginning
goods	technology	given	oil	unit	near	balance	planned	great
management	company	place	carbon	subsidiary	natural	currency	reduce	later
time	data	marketing	cubic	completed	properties	trade	local	remain
costs	network	improved	energy	work	california	exchange	added	reserve
financial	kodak	passed	fuel	dlr	wells	capital	make	equivalent
process	available	addition	million	received	future	foreign	did	imported

TABLE I: Topics and their most probable word lists for Source-LDA, IR-LDA, and CTM.

- 数据集3

  MedlinePlus.

  比较正确数量, JS散度, PMI(Pointwise Mutual Information).

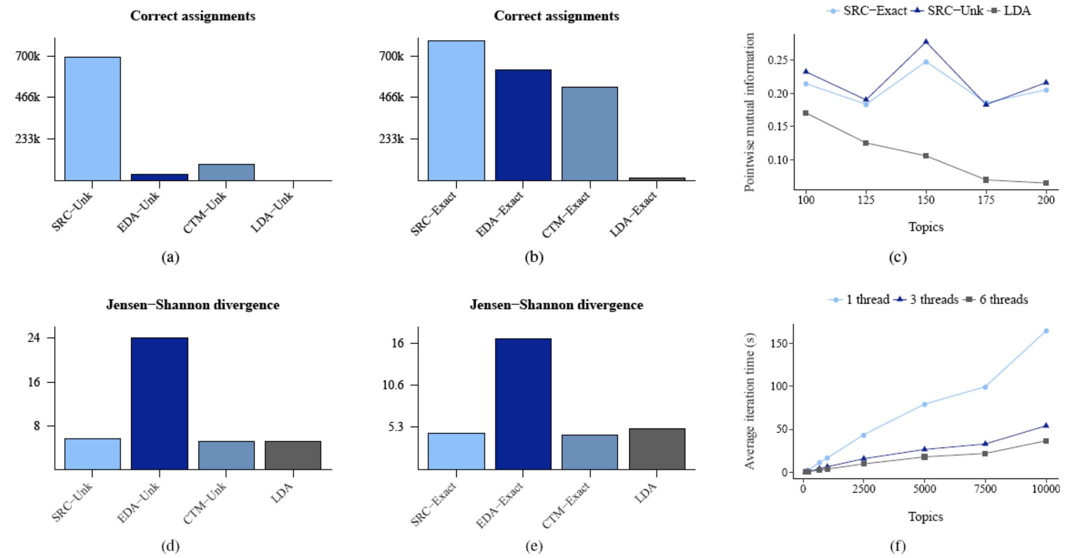


Fig. 8: Results showing the number of correct topic assignments in the mixed model (a) and bijective model (b) and sum total of the JS divergences of θ in the mixed (d) and bijective models (e). Sorted PMI analysis for a Wikipedia generated corpus inferred by the exact bijective model and mixed model is shown by (c). Performance benchmarking is given in (f).

其中SRC-Unk表示上面完整的算法, SRC-Exact表示去掉2,3,4,5行的算法.

- Identifying Rare and Subtle Behaviors: A Weakly Supervised Joint Topic Model, 2011

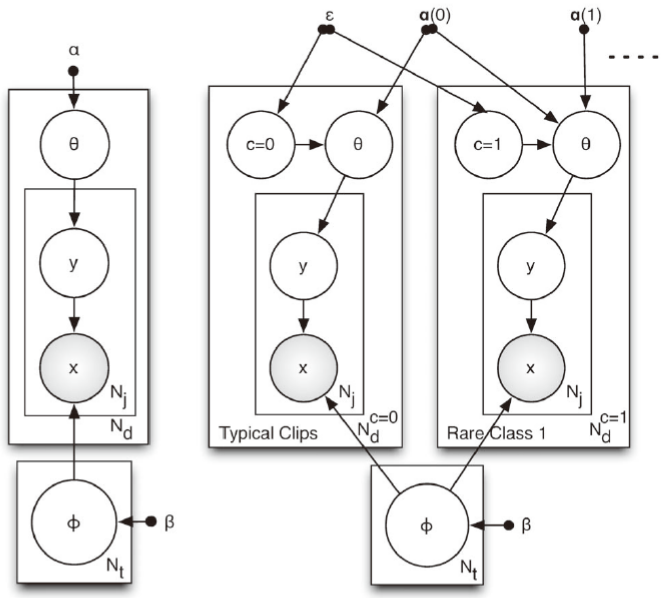
1.问题

识别视频中稀少但重要的场景, 比如交通违法行为.

2.方法

基于 LDA 方法.

将视频数据处理过后形成N_d个文档$X=\{x_j\}_{j=1}^{N_d}$, 每个文档为包含N_j个词的词袋$x_j=\{x_{j,i}\}_{i=1}^{N_j}$. 假设能将文档集分为$N_c+1$个类$X=\{X^c\}_{c=0}^{N_c}$, 第$c$类包含$N_d^c$个文档. 假设其中第0类文档$X^0$只包含典型(typical)主题, $X^c(c>0)$既包含典型主题, 也包含第c类稀有(rare)主题. 设相应的超参数为$\alpha=[\alpha(0),\alpha(1),\cdots,\alpha(N_c)]$, 其中各个参数服从参数为$\epsilon$的多项分布, 则$X^0$由中分布由$\alpha^0=\alpha(0)$生成, $X^c(c>0)$由$\alpha^c=[\alpha(0),\alpha(c)]$生成, 对比 LDA, 有下面的图:



(a) LDA (b) WS-JTM

Fig. 2. (a) LDA [21] and (b) our WS-JTM graphical model structures (only one rare class shown for illustration). Shaded nodes are observed.

算法为:

1. For each activity k, $k = 1, \ldots, N_t$;
   a. Draw a Dirichlet word-activity distribution $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta)$;
2. For each clip j, $j = 1, \ldots, N_d$;
   a. Draw a class label $c_j \sim \mathrm{Multi}(\varepsilon)$;
   b. Choose the shared prior $\boldsymbol{\alpha}^{c=0} \triangleq \boldsymbol{\alpha}(0)$ or $\boldsymbol{\alpha}^{c>0} \triangleq [\boldsymbol{\alpha}(0), \boldsymbol{\alpha}(c)]$.
   c. Draw a Dirichlet class-constrained activity distribution $\boldsymbol{\theta}_j \sim \mathrm{Dir}(\boldsymbol{\alpha}^c)$;
   d. For observed words $i = 1, \ldots, N_w^j$ in clip j:
      i. Draw an activity $y_{j,i} \sim \mathrm{Multi}(\boldsymbol{\theta}_j)$;
      ii. Sample a word $x_{j,i} \sim \mathrm{Multi}(\boldsymbol{\phi}_{y_{j,i}})$.

创新点是作了典型主题与稀有主题服从多项分布的先验假设, 但同时也引入了需要调试的超参数(ϵ). 另外, 还假设某个类中既包含典型主题, 也包含稀有主题, 而前面的Seed-guided DFC似乎假设一个类中只包含一个主题.

3. 实验

- 数据集1

  手动构建的图.

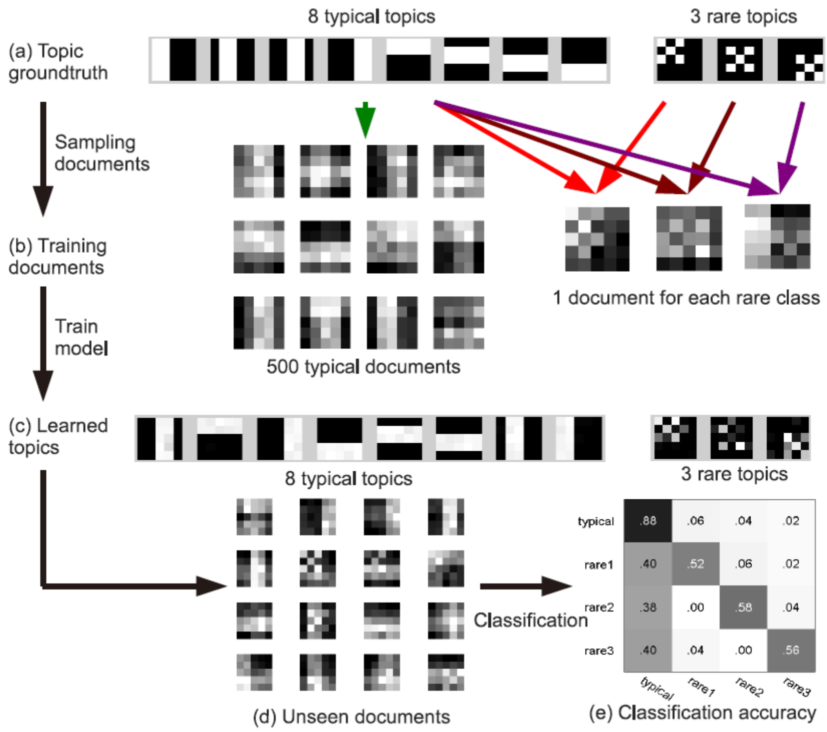


Fig. 3. Illustration and validation of WS-JTM using synthetic data.

控制稀有类中文档数量, 与其他方法对比结果:

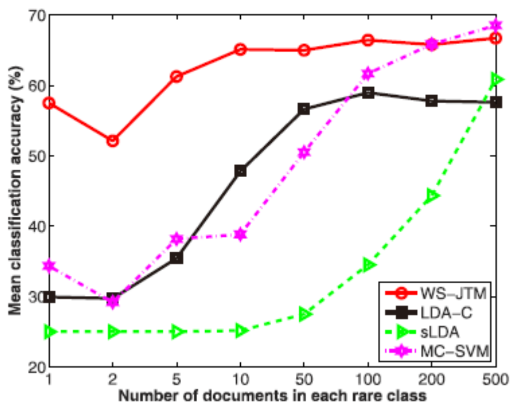


Fig. 4. Synthetic data classification performance as a function of rare-class example sparsity. One-shot learning corresponds to the y-axis. Our WS-JTM exhibits dramatically superior performance in the low data domain.

上图说明稀有类中文档数量较少时本文方法优势明显, 甚至可以减少到1个文档(one-shot learning).

- 数据集2

MIT数据集(30Hz, 720*480, 1.5h)和QMUL数据集(25Hz, 360*288, 1h). 实验中使用的数据分布为:

TABLE 2
Number of Clips Used in the Experiments

MIT			
Total	Typical (300)	Rare 1 (26)	Rare 2 (28)
Train	200	1, 2, 5, 10	1, 2, 5, 10
Test	100	16	18

QMUL			
Total	Typical (200)	Rare 1 (12)	Rare 2 (5)
Train	100	1, 2	1, 2
Test	100	10	3

识别结果为:

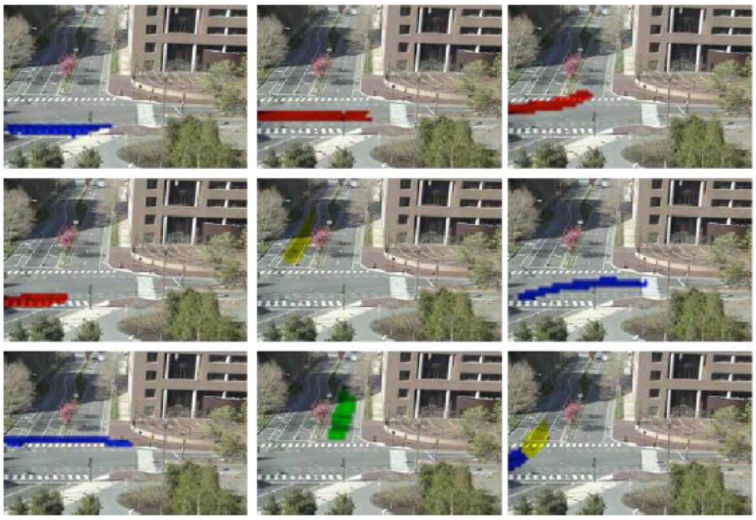


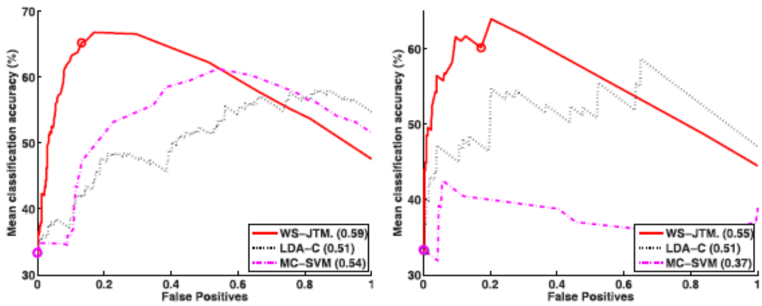
Fig. 6. Typical activities learned from the MIT data set. Red: Right, Blue: Left, Green: Up, Yellow: Down.

(a) Rare 1: Left Turn (b) Rare 2: Right Turn

Fig. 7. One shot learning of rare activities: MIT data set.

QMUL结果图类似, 就不截在这里了.

以上实验都是已知典型主题与稀有主题的分布, 事实上, 我们可能不知道这个分布, 因此会涉及到对ϵ调参. 文章使用了CFAR(constant false alrm rate)方法评估, 此处的"false alarm"指的是被划分为稀有类的典型类, 对比结果为



(a) MIT dataset (b) QMUL dataset

Fig. 13. Average classification accuracy achieved while controlling false alarm rate. Quantity in brackets indicates area under the curve.