关于相似度的主题模型

1、摘要

主题建模已经被广泛用于文本挖掘中。之前的主题模型，例如 LDA，已经能够成功的学习隐藏的主题，但是它们不能利用文本的元数据。为了解决这个问题，许多增强主题模型已经被提出，但是大多数现有的模型只能处理分类和数值类型元数据。我们识别另一种类型的元数据，它可以在某些情景下更自然地获得。这些文档之间的相对相似之处。

在本实验中，我们采用了一个一般模型，它结合了 LDA 与文档相关相似之处衍生的约束。在我们的模型中，限制条件作为 LDA 的对数似然的正则化矩阵。我们使用 Gibbs-EM 来适用提出的模型。采用两个现实世界的数据集进行实验，实验表明我们采用的模型能够学到有意义的话题。该结果还表明，此模型在主题相关及文档分类领域表现优秀。

2、模型

2.1 模型介绍

在相关相似度被给出的情况下，我们来处理文档的主题建模问题。在这部分，我们首先从相关相似度推导制定约束条件。然后简单回顾 LDA 和之前的模型，进而介绍我们采用的模型。

2.2 约束条件

在仅仅知道一些文本的相关相关相似度，而不是全部的情况下，我们来建立文本模型。这种相关相似性可以来自于一个特定的应用程序中人为的判定或从其它数据中自动导出。首先，我们假设给出了一组文档，记为 D。我们进一步假设有距离函数 dist(di,dj),di,dj 来自于 D，来计算相关相似性，我们假设给出一个集合 T 三联体如下：

$$\mathcal{S} = \{(d_i, d_i^+, d_i^-)\}_{i=1}^{T}, \quad (1)$$

where $d_i, d_i^+, d_i^- \in \mathcal{D}$, and d_i is more similar to d_i^+ than to d_i^-. In other words,

$$dist(d_i, d_i^-) > dist(d_i, d_i^+), \quad \forall (d_i, d_i^+, d_i^-) \in \mathcal{S}. \quad (2)$$

受最大边缘方法的启发，我们重写约束条件如下：

$$dist(d_i, d_i^-) \geq dist(d_i, d_i^+) + C, \quad \forall (d_i, d_i^+, d_i^-) \in \mathcal{S}. \quad (3)$$

where C, a positive constant, is a margin to ensure that $dist(d_i, d_i^-)$ is sufficiently larger than $dist(d_i, d_i^+)$.

从上面的定义可知，我们的目标函数为最小化下面的损失函数：

$$\mathcal{L} = \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-), \quad (4)$$

where

$$\mathcal{L}_i(d_i, d_i^+, d_i^-) = \max(0, dist(d_i, d_i^+) + C - dist(d_i, d_i^-)). \quad (5)$$

下面的问题是如何定义距离函数 dist()，距离函数将要基于从 LDA 学习得到的主题分布。

LDA(Latent Dirichlet Allocation)

LDA 是一个广泛使用的主题模型，它起源于 PLSA，LDA 定义了一个特殊的贝叶斯模型，它能够克服 PLSA 中存在的一些问题。

假设文档集为 D，其中的每一个文档 d 包含 Nd 个单词，我们假设存在 K 个主题，每一个与一个多项式单词分布相关。每个文档在 K 维的主题空间中存在一个主题分布。从文档

的主题分布中，文档中的每一个单词拥有一个隐藏的主题标签。LDA 的生成过程被描述为：

- For each topic $k = 1, \ldots, K$, draw $\varphi_k \sim \text{Dir}(\beta)$
- For each document $d \in \mathcal{D}$
  - Draw $\theta_d \sim \text{Dir}(\alpha)$
  - For each word $w_{d,n}, n = 1, \ldots, N_d$
    - Draw $z_{d,n} \sim \text{Multi}(\theta_d)$
    - Draw $w_{d,n} \sim \text{Multi}(\varphi_{z_{d,n}})$

Here α and β are parameters of the Dirichlet priors.

LDA 的参数能过通过几种不同的方法进行学习。

2.3 正规化模型

在这部分，我们介绍我们采用的模型，它结合 LDA 和之前介绍的约束条件，目的是学习更好的主题。

首先，只考虑标准的 LDE 模型，我们的目标是寻找最大化下面所示的目标函数时的最优参数：

$$\log\Big(p(\boldsymbol{w}|\boldsymbol{\theta},\boldsymbol{\varphi})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\varphi}|\boldsymbol{\beta})\Big).$$

为了使这个目标函数与前面介绍的限制条件相联系，我们简单的把这两项相加：

$$\log\Big(p(\boldsymbol{w}|\boldsymbol{\theta},\boldsymbol{\varphi})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\varphi}|\boldsymbol{\beta})\Big) - \eta\sum_{i=1}^{T}\mathcal{L}_i(d_i, d_i^+, d_i^-), \quad (6)$$

where η is a constant to balance the two terms.

现在我们需要定义距离函数 dist，它需要与我们的模型参数相关。直观的，如果如果两个文档拥有相似的 $\boldsymbol{\theta}_d$,它们的距离应该更小。这里有几个选择可以考虑，第一个是欧氏距离，每一个 $\boldsymbol{\theta}_d$ 被当作一个 k 维向量，标准的欧氏距离能够从中计算得到。我们采用均方欧氏距离。另一个选择是 KL-divergence ，定义如下：

$$D_{\text{KL}}(\boldsymbol{\theta}||\boldsymbol{\theta}') = \sum_{k=1}^{K}\boldsymbol{\theta}_k \log\frac{\boldsymbol{\theta}_k}{\boldsymbol{\theta}'_k}.$$

然而，由于 KL-divergence 不是对称的，我们考虑使用对称的 KL-divergence 来代替。

$$\widetilde{dist}(d_i, d_j) = D_{\text{KL}}(\boldsymbol{\theta}_{d_i}||\boldsymbol{\theta}_{d_j}) + D_{\text{KL}}(\boldsymbol{\theta}_{d_j}||\boldsymbol{\theta}_{d_i}).$$

然而，等式 6 是一个约束的优化问题，为了把这个目标函数它转化为一个非限制性优化问题，我们首先定义如下的转化函数：

$$\theta_{d,k} = \frac{e^{\lambda_{d,k}}}{\sum_{k'=1}^{K} e^{\lambda_{d,k'}}}. \tag{7}$$

之后我们改变 Dirichlet prior 为 Gaussian prior，我们把目标函数变化为下面的形式：

$$\mathfrak{L}(\boldsymbol{\lambda}) = \underbrace{\log p(\boldsymbol{w}|\boldsymbol{\lambda}, \boldsymbol{\beta})}_{\text{log likelihood}} + \underbrace{\log p(\boldsymbol{\lambda}|(\mathbf{0}, \sigma^2\mathbf{I}))}_{\text{prior}} - \underbrace{\eta \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-)}_{\text{hinge loss}} \tag{8}$$

Note that the loss $\mathcal{L}_i(d_i, d_i^+, d_i^-)$ is also a function of $\boldsymbol{\lambda}$.

3、Gibbs-EM 法模型拟合

我们采用 Gibbs-EM 方法优化前边提到的目标函数（8），回想一下，我们已经定义隐藏变量ž表示话题任务。 利用隐藏变量，目标函数可以如下来优化。在第 E 步，我们在 tth 迭代修改参数 λ(t) 并获得假定的隐变量分配 p(z|w, $\lambda^{(t)}$, β）。在第 M 步，我们解决了如下的优化问题：

$$\boldsymbol{\lambda}^{(t+1)} = \arg\max \mathbb{E}_{\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\lambda}^{(t)},\boldsymbol{\beta}}[\mathfrak{L}'(\boldsymbol{\lambda})],$$

其中，$E_q[f]$ 是 f 关于分布 q 的预期值，并且

$$\begin{aligned}\mathfrak{L}'(\boldsymbol{\lambda}) &= \log p(\boldsymbol{w}, \boldsymbol{z}|\boldsymbol{\lambda}, \boldsymbol{\beta}) + \log p(\boldsymbol{\lambda}|(\mathbf{0}, \sigma^2\mathbf{I})) \quad (9)\\ &\quad -\eta \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-).\end{aligned}$$

利用 Gibbs-EM，代替评估精确的条件分布 p(z|w, $^{(t)}$, ），我们利用 Gibbs-EM 抽样粗略估计。

3.1 E-步

在 E-step，对于所有的单词，我们通过修改 $\lambda^{(t)}$ 利用 Gibbs 抽样出去隐藏的主题变量 z。为了简化讨论过程，当涉及到主题分布和以及其他时，我们用 θ，它们之间的确定性的关系在等式（7）给出。要执行 Gibbs 抽样，我们需要计算分配一个题目给定的所有其他议题分配给所有的换句话说特定词的概率：

$$p(z_{d,n} = k|\mathbf{w}, \boldsymbol{z}_{-(d,n)}, \boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{p(\boldsymbol{w}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\beta})}{p(\boldsymbol{w}_{-(d,n)}, \boldsymbol{z}_{-(d,n)}|\boldsymbol{\theta}, \boldsymbol{\beta})},$$

其中，(d, n)表明 Zd,n 或 Wd,n 被排除。利用狄利克雷和多项分布的共轭性，Gibbs 更新我们的模型规则可以表示如下：

$$p(z_{d,n} = k|\mathbf{w}, \boldsymbol{z}_{-(d,n)}, \boldsymbol{\theta}, \boldsymbol{\beta}) \propto \frac{n_{k,w_{d,n}} + \beta - 1}{\sum_{v=1}^{V} n_{k,v} + V\beta - 1} \cdot \theta_{d,k}, \quad (10)$$

其中 Nk,v 表示词 v 被分配给主题 k 的次数

正如我们在上一节中指出的那样，我们不直接通过优化目标函数估计。但与 Gibbs 抽样，可估计如下：

$$\hat{\varphi}_{k,v} = \frac{n_{k,v} + \beta}{\sum_{v'=1}^{V} n_{k,v'} + V\beta}. \tag{11}$$

Algorithm 1 Gibbs-EM for our model.

Input:

D documents, # topics K, size of the vocabulary V, regularization parameter η, margin C, max # EM iterations nEM, # Gibbs sampling iterations nGS, max # gradient descent in each M-step nGD.

Output:

$\lambda_{d,k}$ and $\varphi_{k,v}$, $d = 1, .., D; k = 1, .., K; v = 1, .., V$

Randomly initialize $\boldsymbol{z}$ and $\boldsymbol{\lambda}$
$t \leftarrow 0$
while $(t < nEM)$ **do**
  E-step:
  Sample $z_{d,n}$ as in Eqn (10) with nGS iterations
  M-step:
  $n \leftarrow 0$
  while $(n < nGD)$ **do**
    Compute the objective function $\mathfrak{L}'(\boldsymbol{\lambda})$ as in Eqn (9)
    Set the learning rate ξ
    for $(d = 1$ to $D)$ **do**
      for $(k = 1$ to $K)$ **do**
        Compute the partial derivative $\frac{\partial \mathfrak{L}'(\boldsymbol{\lambda})}{\partial \lambda_{d,k}}$
        ${\lambda_{d,k}}^{(t)}_{(n+1)} \leftarrow {\lambda_{d,k}}^{(t)}_{(n)} + \xi \frac{\partial \mathfrak{L}'(\boldsymbol{\lambda})}{\partial \lambda_{d,k}}$
      end for
    end for
    $n \leftarrow n + 1$
  end while
  $t \leftarrow t + 1$
end while
Compute each $\varphi_{k,v}$ as in Eqn (11)

3.2 M-步

在 M 这一步，我们使用之前 E-步获得的最近一次样本 z(t) 和使用梯度下降学习 $\lambda^{(t+1)}$:

$$\boldsymbol{\lambda}^{(t+1)} = \arg\max_{\boldsymbol{\lambda}} \Big(\log p(\boldsymbol{w}, \boldsymbol{z}^{(t+1)} | \boldsymbol{\lambda}, \boldsymbol{\beta}) + \log p(\boldsymbol{\lambda} | (\mathbf{0}, \sigma^2 \mathbf{I})) - \eta \sum_{i=1}^{T} \mathcal{L}_i(d_i, d_i^+, d_i^-) \Big). \tag{12}$$

通过计算公式中的目标函数（12）相对于每个 d,k 的一阶偏导数，我们可以使用梯度下降优化目标函数。算法 1 中总结了模型拟合算法。

4、实验

在本节中，我们给出实验来评价我们的模型。首先，我们描述我们的数据集。然后，我们进行了实验定量和定性。

4.1 数据集

我们使用两个广泛使用的文本语料库，20 新闻组和 TDT2。20 个新闻文本语料库是一个收集约 20000 新闻组文件，分区均匀在 20 个不同的新闻组。我们使用这个版本的数据集预处理，在文件分为训练集和测试集。在训练集的文档中，我们随机选择了 100 个文件，从每个类别和删除停止的话和非常短的文件（文件少于 3 个字），从而留下我们与 14538 个不同的单词的 1997 个文件。

TDT2 语料库包含可分为 96 类的 11,201 文件。只有最大的 20 类别被保留，和那些出现在多个类别的文件被删除。我们也随机抽取 100 个文档从每个类别作为 20 新闻组数据集的策略。最后，还有与 12,166 不同的单词左边的 1,998 文件。

4.2 实验装置

在我们的实验中，我们执行 200 次 Gibbs-EM。每次运行，我们跑了 100 次迭代的吉布斯采样和另一个 10 迭代梯度下降。我们设置狄利克雷先验 = 0.1，高斯先验模型的方差 = 1。我们还固定的主题数是 20 （每个数据集类别的数目相同）。请注意，我们不调整此参数，因为我们的目标不是来找主题的最优数量。

为自动获取模型的三重态约束，我们在训练文档集根据他们的真实类别标签取样一组三重态。具体来说，通过随机抽样两个文件在同一类别和一个文件从另一个不同的类别，我们生成一个三重态实例采用随机抽样。我们的实验中，我们从每个数据集的训练集生成 100 K、 50 K 和 10 K 三重态实例。试验运行在硬件环境为 4 芯和 4GB 内存的机器上。

4.3 定量评价

在这一部分，我们以主题连贯性和文档分类来定量评价我们的模型。我们采用不同数量的采样三重态 （即 100 K、 50 K 和 10 K） 而不是其他基线评估我们的模型。具体来说，在我们的实验中采用的方法是：

- DRS-KL。我们与对称 KL 散度距离度量的模型。
- DRS-SE。我们的模型与平方欧氏距离作为距离的函数。
- LDA。标准的 LDA 模型。
- sLDA。监督的 LDA 模型。

请注意，DRS-KL 和 DRS SE，我们调正则化参数 和边缘 C，并报告结果与最佳的性能。若要测试我们模型的稳定性，我们将五折交叉验证用于所有方法。我们应强调，我们取样三重态情况下仅从训练文档，即没有从测试文档的元数据使用。这将确保我们的模型与基线比较公平。

主题连贯性：

在这个实验中，我们想要比较不同的主题模型的性能。以往的研究通常用作度量的困惑（可能性举行出数据）。然而，这种指标不能测量相干性学习主题。为了解决这个问题，我们探索另一个指标来衡量学习模型的质量。具体来说，要测量主题的语义一致性，我们使用逐点互信息。PMI 度量一些词的共现，被定义为：

$$\mathrm{PMI}(w) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \quad (13)$$

表 1 介绍了所有候选方法的 PMI 分数。从结果中我们可以看到相比 LDA 模型中的所有设置我们的模型实现了更高的 PMI 指数分数。因此我们可以得出结论，我们的模型是在两个语料库中学习主题连贯性更为成功。相比下，LDA 没有展示这样的能力，因为它不捕获文档之间的相对相似的限制。

文档分类：

与话题建模，每个文档可表示其话题分布。在这一小节中，我们研究使用隐藏的主题不同的主题模型文档了解到分类。

我们使用的文档主题分布为特征和训练 RBF 内核支持向量机（SVM）。表 2 示出的分类精度对两个数据集 5 倍交叉验证。

结果表明点数目：

1.与 LDA 相比，sLDA 更好地执行。这是因为 sLDA 利用训练文档的分类信息

2.我们的方法和 SLDA 之间的比较是错杂的，但一般当我们使用多个三元组（100K）和使用对称 K-发散距离，我们的方法是更有可能胜过 SLDA。请注意，我们的方法依赖于相对文档相似性并且 SLDA 需要的文件标签。 此外，在 SLDA，使用所有训练文件的标签，而在我们的模型中，只有一小部分抽样标签被使用。

计算成本：

表 3：训练时间（分钟）比较。

# tps	method	time	
		20 newsgroups	TDT2
100K	DRS-KL	105	47
	DRS-SE	156	76
50K	DRS-KL	77	39
	DRS-SE	89	46
10K	DRS-KL	40	29
	DRS-SE	36	29
–	sLDA	38	58

表 1：基于对两个数据集主题 PMI 主题连贯性。度量越大，主题越好。

		20 newsgroups						TDT2					
# tps*	method	fold					avg*	fold					avg*
		0	1	2	3	4		0	1	2	3	4	
100K	DRS-KL	-1.43	-3.97	-4.75	-4.52	-4.16	-3.77	-0.96	0.36	-1.18	-1.16	-1.17	-0.82
	DRS-SE	-2.41	-2.79	-2.59	-1.28	-2.22	-2.26	0.40	-0.39	-4.09	-0.98	-0.20	-1.05
50K	DRS-KL	-3.41	-1.07	-3.17	-1.63	-4.55	-2.77	-1.76	-1.57	0.21	-1.38	-3.31	-1.56
	DRS-SE	-3.58	-0.44	-1.08	-2.45	-0.84	-1.68	-2.72	0.41	-0.98	-3.12	-3.32	-1.95
10K	DRS-KL	-3.18	-2.59	-4.16	-0.28	-3.41	-2.72	-1.17	0.01	-0.75	-0.19	-2.36	-0.89
	DRS-SE	-3.94	-6.27	-1.44	-5.50	-3.21	-4.07	-3.50	-4.49	-0.79	-0.40	-0.22	-1.88
–	LDA	–	–	–	–	–	-5.71	–	–	–	–	–	-1.98

表 2：两个数据集的分类精度。

		20 newsgroups						TDT2					
# tps	method	fold					avg	fold					avg
		0	1	2	3	4		0	1	2	3	4	
100K	DRS-KL	0.529	0.558	0.642	0.626	0.674	**0.606**	0.824	0.871	0.929	0.926	0.939	**0.898**
	DRS-SE	0.532	0.508	0.518	0.576	0.587	0.544	0.861	0.863	0.908	0.932	0.934	**0.899**
50K	DRS-KL	0.542	0.592	0.605	0.579	0.658	0.595	0.834	0.882	0.932	0.905	0.895	0.889
	DRS-SE	0.555	0.526	0.561	0.547	0.618	0.562	0.845	0.897	0.900	0.926	0.868	0.887
10K	DRS-KL	0.589	0.563	0.537	0.516	0.587	0.558	0.787	0.866	0.911	0.932	0.926	0.884
	DRS-SE	0.568	0.576	0.550	0.539	0.582	0.563	0.863	0.824	0.884	0.932	0.918	0.884
–	LDA	0.518	0.482	0.518	0.539	0.545	0.521	0.784	0.853	0.887	0.871	0.895	0.858
	sLDA	0.518	0.550	0.542	0.582	0.616	0.562	0.837	0.839	0.884	0.932	0.939	0.886

4.4 定性评价

在这一小节中，我们展示通过我们的模型学习到的隐藏主题。对于每个数据集，我们随机选择四个主题并且显示每个主题顶部 10 个单词。表 4 和 5 分别展示出了基于 20 个新闻组数据集和 TDT2 数据集产生的主题单词。

表 4： 20 新闻组模型示例主题。

Topic 1	Topic 2	Topic 3	Topic 4
gun	god	privacy	drive
control	people	encryption	disk
crime	jesus	internet	hard
guns	christian	anonymous	card
rate	bible	information	mb
weapons	man	government	drives
people	religion	email	apple
police	way	technology	system
manes	christ	mail	know
rates	religious	access	dos

表 5: TDT2 数据集示例主题。

Topic 1	Topic 2	Topic 3	Topic 4
iraq	israel	spkr	tobacco
united	israeli	voice	smoking
weapons	palestinian	president	tax
gulf	netanyahu	clinton	industry
iraqi	peace	news	companies
oil	talks	peterjennings	congress
saddam	palestinians	camera	money
gas	arafat	white	billion
war	west	house	settlement
ap	bank	abcnews	bill

我们可以从两个表中发现题目具有一般意义。例如，从最高的字眼“gun,” “control”和 “crime”，它是容易证明表 4 主题 1 是关于"枪支控制";“tobacco”，“smoking”和“tax”等话题字眼"吸烟"和"税"表明，在表 5 中的主题 4 是关于"控烟"。

5、结论

在本文中，我们执行文档之间的相对相似性主题建模。我们制定的约束作为损失函数，并提出与这种限制 LDA 相结合的一般概率模型。我们的模型将限制为文本的数似然的正则化。大量的实验是在两个现实世界的数据集，其中的实证结果表明，我们的模型不仅学习有意义的话题，而且也优于主题连贯性和文档分类任务基线进行广泛的试验。