基于半监督 LDA 的文本分类应用研究

郑世卓，崔晓燕

（北京邮电大学自动化学院，北京 100876）

摘 要：在如今信息数据大爆炸的时代，数据的增长呈现指数级增长，而且其中大部分数据是非结构化数据，这些数据中蕴藏着大量且重要的知识等待着我们用合理的办法将其挖掘出来，如何方便合理快速的进行文本分类也是一个非常重要的课题。LDA 模型是一种无监督的模型，它可以发现隐性的主题，为了更有效的发现隐性主题，本文提出一种基于半监督的 LDA 主题模型，找到一个主题集作为隐性层的知识集，通过这种方法找到的主题与文本更相关，另外，将 LDA 模型与基于半监督 LDA 模型应用于文本的特征提取，并与其它特征提取方法比对，实验表明，半监督 LDA 模型性能略好。

关键字：文本分类；主题模型；LDA 模型；半监督 LDA

中图分类号：TP391.1 **文献标识码：**A DOI：10.3969/j.issn.1003-6970.2014.1.012

本文著录格式：[1] 郑世卓，崔晓燕 . 基于半监督 LDA 的文本分类应用研究 [J]. 软件 ,2014,35(1):46-48

Research on Text Classification Based on Semi-supervised LDA

ZHENG Shi-zhuo, CUI Xiao-yan

（*Automation School, Beijing University of Posts and Telecommunication, Beijing,* 100876）

【Abstract】Nowadays, it’s a time of information explosion and exponential growth of data. Most of the data is unstructured, which bears a large number of important knowledge awaits us to find out. How to work out a convenient and quick way to process text classification is also a very important issue. LDA model is an unsupervised model, which can discover latent topics under unlabeled data. In order to more effectively find latent topics, a model based on semi-supervised LDA is proposed. By finding a topic set to discover more relevant topics. In addition, the LDA model and the semi-supervised LDA model are applied to the text of feature extraction, comparing with other models. The experiments show that semi-supervised LDA model performance slightly better.

【Key words】 text classification; topic model; LDA model; semi-supervised LDA model

0 引言

主题模型 [1] 的应用广泛，一种简单有效的主题标签信息模型被应用到了计算机视觉任务。主题模型方法也被应用到了场景建模，分割及分类或检测。在其中的一些视觉应用中，潜主题 [2] 本身被假定为对应于对象的标签。如果标记的数据是可用的，要么全部或一些 z 值可以被视为可观察到的，而不是潜在的变量。我们的模型将扩展 z 的标记从单值的形式变成子集的形式，从而提供额外的模型表现。

如果文档的基于主题的陈述是用于文档分类，为单词 w 提供 z- 标签 [3] 的话可以被视为作为与特征标签半监督的学习的功能类似。在这里，单词被视为特征，而 z- 标签向导作为一个特征标签。这不同于其他使用文件标签信息的有监督的 LDA 变种，

LDA 模型 [4] 的统计软件调试文件，它只能出现主题在文件的一个特殊的子集。这种作用是通过使用不同的超参数α实现文件的 2 个子集。半监督 LDA 模型可以实现与通过限制在 z 在文件之外的特殊的子集同样的效果，以便在 z 的不能假设主题价值。因此，本方法可以被看作是 LDA[5] 的推广。

另一种看法是，z- 标签可能指导二次发现主题模型或发现不占优势的统计模式数据。用户对这些主题可能更有兴趣或与用户的目标更相关，但标准的 LDA 会忽略掉这些有用的主题，而选取更显著地结构，这便是半监督 LDA 模型的优势所在。

1 模型建立

对于潜在主题z_i，我们假定$C^{(i)}$是可能的 z- 标签的子集，我们通过修改吉布斯采样方程具有指示器的方程$\delta(v \in C^{(i)})$，其取值为 1，如果$v \notin C^{(i)}$，其取值为 0。

我们让

$$q_{iv} = \left(\frac{n_{-i,v}^{(d)} + \alpha}{\sum_{u}^{T}(n_{-i,v}^{(d)} + \alpha)}\right)\left(\frac{n_{-i,v}^{(w_i)} + \beta}{\sum_{w'}^{W}\beta + n_{-i,v}^{(w')}}\right) \quad (1\text{-}1)$$

如上讨论，有如下公式：

$$P(z_i = v \mid z_{-i}, w, \alpha, \beta) \propto q_{iv}\delta(v \in C^{(i)}) \quad (1\text{-}2)$$

如果我们希望限制z_i为单个值，例如让z_i=6，而现在我们就可以通过设置$C^{(i)}$={6}来办到。同样，我们可以通过设置$C^{(i)}$来限制z_i={1，2，3} 到值为 {1，2，3} 的子集。最后，为不受约束的z_i，我们只需设置$C^{(i)}$={1，2，...，T}，在这种情况下，我们修改采样方程为标准的吉布斯抽样方程。

这一提法给了我们一个灵活的方法 [6] 来插入优先级高的知识到潜在主题的推演中。我们可以在语料库中为每一个词w_i设

表 2-2a 有监督的 LDA 模型条件下主题对应的词

Tab. 2-2a corresponding words of LDA model under supervised conditions

主题编号	相关词
主题 0	translation, ribosomal, trna, rrna, initiation, ribosome, protein, ribosomes, is, factor, processing, translational nucleolar, pre-rrna, synthesis, small, 60s, eukaryotic, biogenesis, subunit, trnas, subunits, large, nucleolus factors, 40, synthetase, free, modification, rna, depletion, eif-2, initiator, 40s, ef-3, anticodon, maturation 18s, eif2, mature, eif4e, associated, synthetases, aminoacylation, snornas, assembly, eif4g, elongation

表 2-2b 无监督的 LDA 模型主题对应的词

Tab. 2-2b corresponding words of LDA model without supervised conditions

主题编号	相关词
主题 13	mrna, translation, initiation, mrnas, rna, transcripts, 3, transcript, polya, factor, 5, translational, decay, codon decapping, factors, degradation, end, termination, eukaryotic, polyadenylation, cap, required, efficiency synthesis, show, codons, abundance, rnas, aug, nmd, messenger, turnover, rna-binding, processing, eif2, eif4e eif4g, cf, occurs, pab1p, cleavage, eif5, cerevisiae, major, primary, rapid, tail, efficient, upf1p, eif-2
主题 21	type, is, wild, yeast, trna, synthetase, both, methionine, synthetases, class, trnas, enzyme, whereas, cytoplasmic because, direct, efficiency, presence, modification, aminoacylation, anticodon, either, eukaryotic, between different, specific, discussed, results, similar, some, met, compared, aminoacyl-trna, able, initiator, sam not, free, however, recognition, several, arc1p, fully, same, forms, leads, identical, responsible, found, only, well
主题 43	ribosomal, rrna, protein, is, processing, ribosome, ribosomes, rna, nucleolar, pre-rrna, rnase, small, biogenesis depletion, subunits, 60s, subunit, large, synthesis, maturation, nucleolus, associated, essential, assembly components, translation, involved, rnas, found, component, mature, rp, 40s, accumulation, 18s, 40, particles snornas, factors, precursor, during, primary, rrnas, 35s, has, 21s, specifically, results, ribonucleoprotein, early

置$C^{(i)}$，这让我们，例如，强制两次出现相同的字（例如，”apple pie”和”apple iPod”）通过不同的主题加以解释。这种效果是不可能通过特定主题[7]的非对称β矢量实现并将一些条目设置为零的。

这种硬约束模型也可以适当放宽，让$0 \le \eta \le 1$成为我们的约束的另一个条件，其中当$\eta = 1$时，就又成为了硬约束，而当$\eta = 0$时，就成为了无约束的采样方程。

$$P(z_i = v | z_{-i}, w, \alpha, \beta) \propto q_{iv}(\eta\delta(v \in C^{(i)}) + 1 - \eta) \quad (1\text{-}3)$$

而我们给出的 z- 标签约束作为机械的修改吉布斯抽样方程[8]，它可以从一个无向延伸衍生出 LDA 模型[9]。该软约束吉布斯抽样方程是由这种模型自然产生，它也是后面描述的一阶逻辑的约束条件的基础。

2 实验分析

2.1 半监督 LDA 性能评估

实验英文语料库选取 newsgroup，训练集占 70%，测试集占 30%，表中显示了语料库的一部分类别分布。参数设置为：主题数 K 取 50，α和β取值根据经验分别取的是α为 1，β为 0.1，迭代次数为 2000 轮，对于半监督 LDA，将γ取值为 0，每个类别指定一个确定的主题。

表 2-1 各类文本在语料库中的分布

Tab. 2-1 Distribution of various types of texts in the corpus

类别编号	类别	文本数量	文本分布
1	comp.graphics	973	19.89%
2	comp.os.ms-windows.misc	985	20.14%
3	comp.sys.ibm.pc.hardware	982	20.28%
4	comp.sys.mac.hardware	963	19.69%
5	comp.windows.x	988	20.20%

现在用实验结果来展示主题集模型，除非另有规定，分别取对称超参数α =0.5，β = 0.1。迭代次数选择 2000 轮，同时所有的 MCMC 链都对 2000 个样品进行估算。

我们探索运用主题集的识别相关的一个目标概念的话，给定的一组与此概念相关的种子单词集。例如，如果生物专家可能会对“translation”的概念感兴趣。接着专家将随后提供一组与这一概念关联很强的相关种子单词集，在这里我们假设种子词集 {translation，trna，anticodon，ribosome}。我们为所有出现这四个字的语料库与添加硬约束z_i =0。然后，运行 LDA，主题数 T 选择 50，对种子单词集，我们分别对标准 LDA 模型和有监督的 LDA 模型进行执行。表 2 给出了无监督的 LDA 模型和有监督的 LDA 模型，运行后的结果。表 2-2a 所示的是有监督的

LDA 模型执行后，与种子单词最相关的 50 个单词的结果，表 2-2b 所示的是在无监督情况下，与种子单词最相关的 50 个单词的结果。

为了更好地理解表中的结果，种子单词用蓝色表示，与种子单词相关的单词用红色表示出来，黑色单词表示与种子单词不相关，从整体上的效果来看，有如下结果：

1）LDA 模型在有监督的情况下，主题 0 和主题 13 所示，有监督模型得到的单词数更多；

2）在与种子单词的相关单词数量上几乎相同的情况下，如主题 21 所示，所指向的主题有所偏移，标准的 LDA 模型更指向 mRNA，而不是指向 translation，存在类似的结果，主题 43 更指向 ribosome 而不是翻译的过程。

这些结论表明，有监督的 LDA 模型有着更加均衡，更加全面的效果。

2.2 LDA 应用实验分析

实验采用 NEWSGROUP 数据集，采用 70% 数据作为训练集，剩余的 30% 数据作为测试集。分类器采用线性 SVM 分类器，通过 Liblinear 工具实现。实验先在训练集上进行 LDA 训练，其中 LDA 训练采用 Gibbs 抽样方法，参数设置为：主题数 K 取 50，α和β的取值根据经验分别取的是α为 1，β为 0.1，迭代次数为 2000 轮，对于半监督 LDA，将γ取值为 0，每个类别指定一个确定的主题。

文本表示选择用 VSM，特征选择分别采用 MI，IG，DF，CHI，LDA，半监督 LDA 进行对比实验，对于半监督 LDA，由于主题的指定，这使得对于种子文档进行吉布斯采样时，种子文档中的词也就属于指定的主题，进行吉布斯采样的文档集包含训练文档和测试文档。特征词的词向量对于半监督 LDA 来说并没有选取所有主题下的值，而是选取那些类别指定的主题值，因为这些主题值有着更加相关的信息，这意味着词向量只选取了 50 维特征向量中的一部分，这里我们选取了一半 25 维的特征。

表 2-3 文本分类中特征选择的结果对比

Tab. 2-3 Results of Feature selection in text categorization

	0.2	0.4	0.6	0.8	1.0
MI	42.883	75.293	78.819	79.820	85.823
IG	78.098	83.839	85.301	85.623	85.823
DF	83.690	85.376	85.763	85.782	85.823
CHI	83.522	85.012	85.523	85.873	85.823
LDA	82.683	85.323	85.633	85.754	85.823

由表 2-3 可以看出，将 LDA 应用于特征选择的效果明显较好，与其他方法相比，比其中最好的特征选择方法 CHI 略好，又由于 LDA 是一种无监督的特征选择方法，所以 LDA 应用于特征选择是一个不错的选择。

表 2-4 半监督 LDA 在文本分类中特征选择的结果对比

Tab. 2-4 Results of Feature selection in text categorization with semi-supervised LDA

	0.2	0.4	0.6	0.8	1.0
MI	42.883	75.293	78.819	79.820	85.823
IG	78.098	83.839	85.301	85.623	85.823
DF	83.690	85.376	85.763	85.782	85.823
CHI	83.522	85.012	85.523	85.873	85.823
半监督 LDA	83.114	85.680	85.732	85.764	85.823

从上表 2-4 可以看出，利用半监督 LDA 的指定主题下的词向量来进行特征选择的效果要比其他特征选择略好，与用原始 LDA 的词向量的效果相差不多，略好，原因可能在于这里面仅使用了指定类别对应的主题信息，而将其他忽略掉的主题可能有些影响较大的特征，所以带来效果的提升不够明显。

3 结论

本文给出了一种半监督的 LDA 模型，这种 LDA 模型将隐性层的主题 z 显示的进行监督，对 Gibbs 采样进行了改进，加入了监督因子，并且利用 LDA 模型进行向量模型表示，看其对文本的表示效果，并其他几种模型相比，对 LDA 模型进行了特征选择的实验分析，其分析结果表明特征选择的表现良好，是一种良好的特征选择方法，基于半监督的 LDA 模型与原始 LDA 模型在特征选择上性能相差不多。

参考文献

[1] David M B, Andrew Y N, Michael I J. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research[J],2003, 3:993-1022.

[2] Jordan B, David B, Zhu X J. 2007. A topic model for word sense disambiguation. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning[J]. 2007, 3:1024-1033.

[3] Gregory D, Gideon M, Andrew M. Learning from labeled features using generalized expectation criteria. SIGIR[J] 2008: 595-602.

[4] Thomas G, Mark S. Finding scientific topics. Proceedings of the National Academy of Sciences[J], 2009, 101(1):5228-5235.

[5] Erik T S, Fien D M. Language independent named entity recognition. In Proceedings of CoNLL-2003[J], 2003:142-147.

[6] Liu Ying, Ciliax B J, Borges K, et al. Comparison of Two Schemes for Automatic Keyword Extraction from MEDLINE for Functional Gene Clustering[C]llProc. of IEEE Computational Systems Bioinformatics Conference. Stanford, Califomia, USA: IEEE Press, 2004: 394-404.

[7] Shi Jing,Hu Ming, Shi Xin, et al. Text Segmentation Based on Model LDA[J]. Chinese Joumal of Computers, 2008, 31 (10):1865-1873

[8] 魏钰洁，潘清，田园 . 基于 IAE 的国防采办系统研究与设计 [J]. 软件，2012,33（9）:17-21

[9] 马华，王清，韩忠东，张西学，郝刚 . 决策树分类算法在个性化图书推荐中的应用 [J]. 软件，2012,33（8）:100-101