Tr-SLDA:一种面向交叉领域的迁移主题模型

唐焕玲[1,2,4],郑 涵[2],刘艳红[1],马思源[2],窦全胜[1,3,4],鲁明羽[5]

(1. 山东工商学院计算机科学与技术学院,山东烟台 264005; 2. 山东工商学院信息与电子工程学院,山东烟台 264005;
3. 山东省高等学校协同创新中心:未来智能计算,山东烟台 264005;
4. 山东省高校智能信息处理重点实验室(山东工商学院),山东烟台 264005;
5. 大连海事大学信息科学技术学院,辽宁大连 116026)

摘 要: 当目标领域缺少足够多的标注数据时,迁移学习利用相关源领域的标注数据,辅助提升目标域的学习性能,但是目标域与源域的数据通常不满足独立同分布,容易导致"负迁移"问题. 本文在有监督主题模型(Supervised LDA,SLDA)的基础上,融合迁移学习方法提出一种共享主题知识的迁移主题模型(Transfer SLDA,Tr-SLDA),提出 Tr-SLDA-Gibbs 主题采样新方法,在类别标签的约束下对不同领域文档中的词采取不同的采样策略,且无需指定主题个数. 辅助源域与目标域共享潜在主题空间,Tr-SLDA 通过发现潜在共享主题与不同领域类别之间的语义关联从源域迁移知识,可以有效解决"负迁移"问题. 基于 Tr-SLDA 迁移主题模型提出 Tr-SLDA-TC(Tr-SLDA Text Categorization)文本分类方法. 对比实验表明,该方法可有效利用源域知识来提高目标领域的分类性能.

关键词: 文本分类; 主题模型; 吉布斯采样; 迁移学习; 负迁移

中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2021)03-0605-09

电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20200210

Tr-SLDA:A Transfer Topic Model for Cross-Domains

TANG Huan-ling[1,2,4],ZHENG Han[2],LIU Yan-hong[1],
MA Si-yuan[2],DOU Quan-sheng[1,3,4],LU Ming-yu[5]

(1. *College of Computer Science and Technology,Shandong Technology and Business University,Yantai,Shandong* 264005,*China*;
2. *College of Information and Electronic Engineering,Shandong Technology and Business University,Yantai,Shandong* 264005,*China*;
3. *Co-innovation Center of Shandong Colleges and Universities:Future Intelligent Computing,Yantai,Shandong* 264005,*China*;
4. *Key Laboratory of Intelligent Information Processing in Universities of Shandong(Shandong Technology and Business University),Yantai,Shandong* 264005,*China*; 5. *Information Science and Technology College,Dalian Maritime University,Dalian,Liaoning* 116026,*China*)

Abstract: With enough labeled data lacking in the target domain,it works well for transfer learning to use the labeled data of the related source domain and help improve the learning performance of the target domain. However,the data of these two domains usually do not satisfy the independently identically distribution,which easily leads to the problem of "negative transfer". Tr-SLDA(Transfer SLDA),a transfer topic model based on supervised topic model (Supervised LDA,SLDA) is proposed,which shares topic knowledge by integrating transfer learning. A Tr-SLDA-Gibbs sampling method is proposed,under the constraints of category labels,different sampling strategies are adopted for words in the documents of different domains without specifying the number of topics. The source domain and target domain share the potential topic space,Tr-SLDA can effectively solve the problem of "negative transfer" by discovering the semantic correlation between the potential shared topics and categories of different domains. The Tr-SLDA-TC (Tr-SLDA-Text Categorization) text classification method is proposed based on the Tr-SLDA model. The comprehensive experiments show that the proposed method can effectively improve the performance of the classification by utilizing the knowledge from the source domain.

Key words: text categorization;topic model;Gibbs sampling;transfer learning;negative transfer

1 引言

传统文本分类模型的建立,需要满足两个基本假设[1]:(1)测试样本与训练样本服从独立同分布假设;(2)有足够可利用的已标注训练样本. 但在实际应用中通常难以满足这两种假设,许多学者运用迁移学习

收稿日期:2020-02-30;修回日期:2020-04-21;责任编辑:梅志强
基金项目:国家自然科学基金(No. 61976124,No. 61976125,No. 61773244,No. 61772319,No. 61873177,No. 61972235);山东省高校科研计划(No. J18KA385),烟台市重点研发计划(No. 2017ZH065,No. 2019XDHZ081)

(transfer learning)技术,从相关源领域中迁移知识,可以较好地解决目标领域的学习任务. 文献[2]提出了一种结合迁移学习的协同过滤方法(Transfer by Collective Factorization,TCF),通过迁移辅助数据中的二进制偏好数据来降低目标域评级数据稀疏性对模型的影响,方法构建了一个共享的潜在空间,并分别学习数据的依存效应. 文献[3]提出一种图共正则化迁移学习(Graph Co-Regularized Transfer Learning,GTL)框架,方法通过保留各域间的统计特性来提取潜在共享知识,同时保留各域间的几何结构来细化潜在知识. 文献[4]提出一种层次迁移注意网络(Transfer Hierarchical Attention Network,THAN)来学习生成式对话系统中的上下文表示,THAN 利用关键词提取和句子蕴涵(sentence entailment)两个相关任务提取句法结构和语义关系,来增强对话系统的词级和句级注意机制. 迁移学习广泛应用于各研究领域,如推荐系统[5],古诗词情感分析[6],案例分类[7]等. 但由于目标领域和源领域不满足同分布假设,上述方法不能有效解决迁移学习中存在的"负迁移[1]"问题是迁移学习需要解决的难点问题.

主题模型(topic model)是篇章级文本语义理解的重要工具,它善于从一组文档中抽取若干组关键词来表达该文档集的核心思想. 典型的是 Blei 等人提出的 LDA(Latent Dirichlet Allocation)主题模型[8],用词的概率分布来表示主题,通过隐含主题来建立语义相近词之间的关联,将文本从高维的词表示变换到低维的主题表示. 主题模型在许多应用领域取得比较多的研究成果,如聚类和分类[9~12]、词义消歧[13]、情感分析[14]和图像处理领域的目标发现与定位[15]、图像分割[16]、图像标注[17]等任务. 文献[18]基于 LDA 主题模型提出一种 SLDA 有监督主题模型,在 LDA 主题模型的基础上引入了用以表示主题-类别分布的新参数,提出新的采样方法,建立主题与类别之间的精准映射,提高了主题模型在文本分类任务上的分类性能. 文献[19]基于分层狄利克雷过程提出一种半监督 HDP 主题模型(Semi-supervised Labeled HDP,SLHDP),应用于标注数据较少,但存在大量未标注数据的半监督文本分类任务. 然而,融合迁移学习的迁移主题模型相关研究尚不多见.

针对目标领域仅有较少标注数据,相关辅助领域存在大量标注数据,二者标注数据关联的类别空间不同,但共享部分类别标签的文本分类任务,传统迁移学习方法难以解决"负迁移"的问题. 本文在 SLDA 有监督主题模型的基础上,融合迁移学习技术提出一种共享主题知识的 Tr-SLDA 迁移主题模型. (1)引入两个新参数 $\boldsymbol{\delta}$ 和 $\boldsymbol{\mu}$,分别表示共享主题与目标领域和源领域的类别之间的概率分布;(2)依据不同领域的类别约束,提出新的 Tr-SLDA-Gibbs 主题采样和参数估计方法,旨在通过发现潜在的共享主题与不同领域的类别之间的语义关联,从而能够有效解决"负迁移"问题;(3)基于迁移主题模型,提出 Tr-SLDA-TC 文本分类方法.

2 Tr-SLDA 迁移主题模型

2.1 Tr-SLDA 概率图模型

SLDA 概率图模型如图 1 所示,$D=\{(\boldsymbol{w}_m,y_m)\}_{m=1}^{M}$ 表示训练文档集,$y_m\in[1,C]$ 表示第 m 篇文档的类别标号,$\boldsymbol{w}_m$ 是第 m 篇文档的词袋向量,N_m 表示第 m 篇文档的长度,C 是类别总数,M 是文档总数,K 是主题数,$w_{m,n}$ 表示第 m 篇文档中第 n 个词,$z_{m,n}$ 是分配给 $w_{m,n}$ 的主题. SLDA 模型有三个参数 $\boldsymbol{\theta}_m$、$\boldsymbol{\varphi}_k$ 和 $\boldsymbol{\delta}_k$,其中 $\boldsymbol{\theta}_m$ 为文档 m 的主题概率分布,$\boldsymbol{\varphi}_k$ 表示第 k 个主题的词概率分布,$\boldsymbol{\delta}_k$ 表示第 k 个主题的类别概率分布. $\boldsymbol{\theta}_m$、$\boldsymbol{\varphi}_k$ 和 $\boldsymbol{\delta}_k$ 服从 Dirichlet 分布,$\boldsymbol{\alpha}$、$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$ 是 Dirichlet 分布的先验参数.

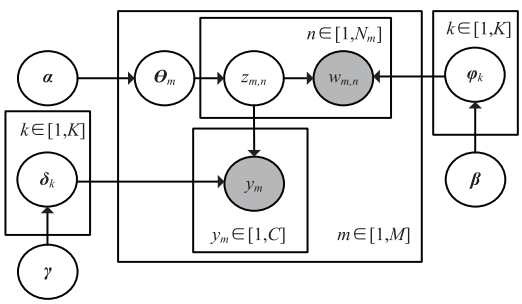


图1 SLDA概率图模型

Tr-SLDA 主题模型的概率图模型如图 2 所示,训练文档集 $D=D^{s}\cup D^{t}$,其中目标领域的文档集 $D^{t}=\{(\boldsymbol{w}_m,y_m)\}_{m=1}^{M_t}$,$y_m\in\Lambda^{t}=\{1^{t},\cdots,C^{t}\}$,$\Lambda^{t}$ 是 D^{t} 的类标签集合(如圆角虚线矩形);辅助源领域的文档集 $D^{s}=\{(\boldsymbol{w}_m,y_m)\}_{m=M_t+1}^{M}$,$y_m\in\Lambda^{s}=\{1^{s},\cdots,C^{s}\}$,$\Lambda^{s}$ 是 D^{s} 的类标签集合(如直角虚线矩形),二者拥有不同的类别空间,$\Lambda^{ts}=\Lambda^{t}\cap\Lambda^{s}$ 是 D^{t} 和 D^{s} 共享类标签集合(如直角与圆角虚线矩形交叉部分). M 是总文档数. D^{t} 和 D^{s} 共享潜在主题空间为 $T=\{1,\cdots,K\}$. 阴影表示的 $w_{m,n}$ 和 y_m 是可观察的,$z_{m,n}$ 是分配给 $w_{m,n}$ 的隐含主题. Tr-SLDA 主题模型的参数为 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}$、$\boldsymbol{\delta}$ 和 $\boldsymbol{\mu}$. 其中 $\boldsymbol{\theta}$ 表示文档-主题分布,$\boldsymbol{\varphi}$ 表示主题-词分布,$\boldsymbol{\delta}$ 表示主题-Λ^{t} 类之间的隐含语义分布,$\boldsymbol{\mu}$ 表示主题-Λ^{s} 类之间的隐含语义分布. $\boldsymbol{\theta}$、$\boldsymbol{\varphi}$、$\boldsymbol{\delta}$ 和 $\boldsymbol{\mu}$ 服从 Dirichlet 分布,$\boldsymbol{\alpha}$、$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$ 和 $\boldsymbol{\eta}$ 是相应的 Dirichlet 分布的先验参数.

2.2 Tr-SLDA 参数估计

Tr-SLDA 迁移主题模型需要估计的参数有 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}$、$\boldsymbol{\delta}$ 和 $\boldsymbol{\mu}$. 在 SLDA-Gibbs 采样方法[18]的基础上提出一种 Tr-SLDA-Gibbs 隐含主题采样方法,其思想是:对 D^{t} 和 D^{s} 中的文档依据不同类别空间的类别约束,采用不同的隐含主题采样策略. 在对每个文档的每个词分配了隐含

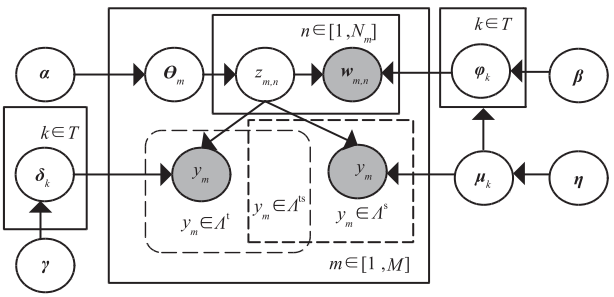


图2　Tr-SLDA主题模型概率图模型

主题后,Tr-SLDA 主题模型的参数 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}$、$\boldsymbol{\delta}$ 和 $\boldsymbol{\mu}$ 可以通过统计频次计算得到,受文献[18]启发,模型最终将得到类别与主题之间的精准映射,所以本文在采样初始化时,为不同类别文档中的词赋以不同的主题编号作为先验,可有效避免"无用主题"[18]的产生,且无需设定主题数. 具体采样步骤如下:

Step1　将不同类别文档中的词初始化为不同的主题编号.

Step2　对 D^{t} 和 D^{s} 中的每个标注文档的每个单词,分别依据文档的类别约束进行隐含主题采样,文档 m 的第 i 个单词 t 的隐含主题计算如下:

$$p(z_i=k\mid \boldsymbol{z}_{-i},\boldsymbol{y}_{-m},\boldsymbol{w})\propto$$

$$\begin{cases}\dfrac{n_{k,-i}^{(t)}+\beta_t}{\sum\limits_{v}^{V}n_k^{(v)}+\beta_v}\cdot(n_{m,-i}^{(k)}+\alpha_k)\cdot(n_{k,-m}^{(j^{\mathrm{t}})}+\gamma_{j^{\mathrm{t}}}),\ \text{if } y_m=j^{\mathrm{t}}\in\Lambda^{\mathrm{t}}\\ \dfrac{n_{k,-i}^{(t)}+\beta_t}{\sum\limits_{v}^{V}n_k^{(v)}+\beta_v}\cdot(n_{m,-i}^{(k)}+\alpha_k)\cdot(n_{k,-m}^{(j^{\mathrm{s}})}+\mu_{j^{\mathrm{s}}}),\ \text{if } y_m=j^{\mathrm{s}}\in\Lambda^{\mathrm{s}}\\ \dfrac{n_{k,-i}^{(t)}+\beta_t}{\sum\limits_{v}^{V}n_k^{(v)}+\beta_v}\cdot(n_{m,-i}^{(k)}+\alpha_k)\cdot(n_{k,-m}^{(j^{*})}+\gamma_{j^{\mathrm{t}}})\\ \quad\cdot(n_{k,-m}^{(j^{*})}+\mu_{j^{\mathrm{s}}}),\ \text{if } y_m=j^{\mathrm{t}}=j^{\mathrm{s}}\in\Lambda^{\mathrm{ts}}\end{cases}\tag{1}$$

其中,V 表示数据集词典的长度,$\boldsymbol{z}_{-i}$ 表示剔除向量 $\boldsymbol{z}$ 的第 i 项,$\boldsymbol{y}_{-m}$ 表示剔除向量 $\boldsymbol{y}$ 的第 m 项,$n_{k,-i}^{(t)}$ 表示剔除 $\boldsymbol{z}$ 的第 i 项(即第 i 个词 $w_i=t$)后,主题 k 分配给词 t 的次数,β_t 表示单词 t 的 Dirichlet 先验. $n_{m,-i}^{(k)}$ 表示剔除 $\boldsymbol{z}$ 的第 i 项后,主题 k 分配给文档 m 中单词的次数,α_k 表示主题 k 的 Dirichlet 先验.

根据文档的类别,约束采样规则如下:

(a)若 $y_m=j^{\mathrm{t}}\in\Lambda^{\mathrm{t}}$,则除了文档 m,从 D^{t} 的其它类别为 j^{t} 文档中进行隐含主题采样,$n_{k,-m}^{(j^{\mathrm{t}})}$ 是剔除文档 m,主题 k 分配为类别 j^{t} 的 D^{t} 单词数;

(b)若 $y_m=j^{\mathrm{s}}\in\Lambda^{\mathrm{s}}$,则除了文档 m,从 D^{s} 的其它类别为 j^{s} 文档中进行隐含主题采样,$n_{k,-m}^{(j^{\mathrm{s}})}$ 是剔除文档 m,主题 k 分配为类别 j^{s} 的 D^{s} 单词数;

(c)若 $y_m=j^{\mathrm{t}}=j^{\mathrm{s}}\in\Lambda^{\mathrm{ts}}$,则除了文档 m,从 D^{t} 和 D^{s} 的其它类别分别为 j^{t} 和 j^{s} 文档中进行隐含主题采样.

Step3　转入 Step2,直到重复迭代 T 次,或者模型的困惑度(perplexity)收敛,perplexity 计算如下:

$$\text{perplexity}=\exp\left\{-\frac{\sum\limits_{m=1}^{M}\log p(w_m)}{\sum\limits_{m=1}^{M}N_m}\right\}\tag{2}$$

$$\begin{aligned}p(\boldsymbol{w}_m)&=\prod_{n=1}^{N_m}\sum_{k=1}^{K}p(z_n=k\mid\boldsymbol{w}_m)\cdot p(w_{m,n}\mid z_n=k)\\&=\prod_{v=1}^{V}\left(\sum_{k=1}^{K}\varphi_{k,v}\cdot\theta_{m,k}\right)^{n_m^{(v)}}\end{aligned}\tag{3}$$

其中,$n_m^{(v)}$ 表示单词 v 在文档 m 中出现的次数.

Step4　计算模型参数 $\boldsymbol{\theta}$、$\boldsymbol{\varphi}$、$\boldsymbol{\delta}$ 和 $\boldsymbol{\mu}$.

$$\theta_{m,k}=(n_m^{(k)}+\alpha_k)\Big/\left(\sum_{z\in T}n_m^{(z)}+\alpha_z\right)\tag{4}$$

$$\varphi_{k,t}=(n_k^{(t)}+\beta_t)\Big/\left(\sum_{v=1}^{V}n_v^{(t)}+\beta_v\right)\tag{5}$$

$$\delta_{k,j}=(n_k^{(j)}+\gamma_j)\Big/\left(\sum_{c\in\Lambda^{\mathrm{t}}}n_k^{(c)}+\gamma_c\right),\ j=j^{\mathrm{s}}\in\Lambda^{\mathrm{t}}\tag{6}$$

$$\mu_{k,j}=(n_k^{(j)}+\eta_j)\Big/\left(\sum_{c\in\Lambda^{\mathrm{s}}}n_k^{(c)}+\eta_c\right),\ j=j^{\mathrm{t}}\in\Lambda^{\mathrm{s}}\tag{7}$$

其中,$m=\{1\cdots M\}$,$k\in T$,$t=\{1\cdots V\}$,$\theta_{m,k}$ 表示文档 m 分配为共享主题 k 的概率,$\varphi_{k,t}$ 表示共享主题 k 分配给词 t 的概率,$\delta_{k,j}$ 表示共享主题 k 属于类别 $j\in\Lambda^{\mathrm{t}}$ 的概率,$\mu_{k,j}$ 表示共享主题 k 属于类别 $j\in\Lambda^{\mathrm{s}}$ 的概率. $n_m^{(k)}$ 表示文档 m 分配给主题 k 的次数,$n_k^{(t)}$ 表示主题 k 分配给单词 t 的次数,$n_k^{(j)}$ 表示主题是 k 分配为类别 j 的单词数.

对比 LDA 采样公式,式(1)引入类别信息 $\boldsymbol{y}_{-m}$,表示模型在类别标签的约束下进行文档词的隐含主题采样,因为理论上相同类别的文档其主题分布是相似的[18]. 而对比 SLDA 采样公式,式(1)对来自不同域文档中的词分别采取不同的采样策略,通过发现潜在的共享主题与不同领域的类别间的语义关联,可有效避免目标域数据受源域数据的影响产生主题偏移和"负迁移".

2.3　Tr-SLDA 测试文档主题推断

测试文档 x 的主题空间是 D^{t} 和 D^{s} 共享主题空间 T,对 x 的词 $\hat{w}_i=e$ 的隐含主题 $\hat{z}_i$ 进行采样,计算如下:

$$\begin{aligned}&p(\hat{z}_i=k\mid\hat{w}_i=e,\hat{\boldsymbol{z}}_{-i},\hat{\boldsymbol{w}}_{-i})\\&\propto\frac{n_{k,-i}^{(e)}+\hat{n}_{k,-i}^{(e)}+\beta_t}{\sum\limits_{v=1}^{V}n_k^{(v)}+\hat{n}_k^{(v)}+\beta_v}\cdot(\hat{n}_{x,-i}^{(k)}+\alpha_{k'})\end{aligned}\tag{8}$$

其中 $\hat{n}_{k,-i}^{(e)}$ 表示新文档剔除第 i 项后,主题 k 分配给词 e 的次数,$\hat{n}_{x,-i}^{(k)}$ 表示剔除第 i 项后,文档 x 分配给主题 k 的次数,其余符号含义同上.

文档 x 属于第 k 个隐含主题的概率 $\hat{\theta}_{x,k}$ 计算如下：

$$\hat{\theta}_{x,k} = \frac{\hat{n}_x^{(k)} + \alpha_k}{\sum_{r\in T}\hat{n}_x^{(r)} + \alpha_r} \tag{9}$$

3 基于 Tr-SLDA 文本分类

基于训练完成的 Tr-SLDA 主题模型，令测试文档 x 的向量为 $\hat{\boldsymbol{w}}$，对新文档 x 的预测类别 $\hat{y}$，提出计算方法：

$$p(\hat{y}|\hat{\boldsymbol{z}},\hat{\boldsymbol{w}}) \propto \begin{cases} g(j^{\mathrm{t}}), & \text{if } \hat{y}=j^{\mathrm{t}},\ j^{\mathrm{t}}\in\Lambda^{\mathrm{t}} \\ \lambda g(j^{\mathrm{s}}), & \text{if } \hat{y}=j^{\mathrm{s}},\ j^{\mathrm{s}}\in\Lambda^{\mathrm{s}} \\ g(j^{\mathrm{t}})+\lambda g(j^{\mathrm{s}}), & \text{if } \hat{y}=j^{\mathrm{t}}=j^{\mathrm{s}},\ j^{\mathrm{t}}\in\Lambda^{\mathrm{ts}} \end{cases} \tag{10}$$

其中：$g(j^{\mathrm{t}}) = p(\hat{y}=j^{\mathrm{t}}|\hat{\boldsymbol{z}}=k)\cdot p(\hat{\boldsymbol{z}}=k|\hat{\boldsymbol{w}})$，$g(j^{\mathrm{s}}) = p(\hat{y}=j^{\mathrm{s}}|\hat{\boldsymbol{z}}=k)\cdot p(\hat{\boldsymbol{z}}=k|\hat{\boldsymbol{w}})$，$\lambda\in[0,1]$ 为平衡因子.

由 Tr-SLDA 模型可知，D^{t} 的隐含主题-类别分布 $p(\hat{y}=j^{\mathrm{t}}|\hat{\boldsymbol{z}}=k)$ 由参数 $\boldsymbol{\delta}$ 揭示，而 D^{s} 的隐含主题-类别分布 $p(\hat{y}=j^{\mathrm{s}}|\hat{\boldsymbol{z}}=k)$ 由参数 $\boldsymbol{\mu}$ 揭示，$p(\hat{\boldsymbol{z}}|\hat{\boldsymbol{w}})$ 则由测试文档 x 的主题概率分布 $\hat{\boldsymbol{\theta}}_x$ 表示. 故，对新文档 x 的类别 $\hat{y}$ 的预测如下：

$$\hat{y} = \arg\max_{\hat{y}=j^{\mathrm{t}}\text{ or }j^{\mathrm{s}}} \begin{cases} \delta_{k,j^{\mathrm{t}}}\hat{\theta}_{x,k}, & \text{if } \hat{y}=j^{\mathrm{t}},\ j^{\mathrm{t}}\in\Lambda^{\mathrm{t}} \\ \lambda\mu_{k,j^{\mathrm{s}}}\hat{\theta}_{x,k}, & \text{if } \hat{y}=j^{\mathrm{s}},\ j^{\mathrm{s}}\in\Lambda^{\mathrm{s}} \\ \delta_{k,j^{\mathrm{t}}}\hat{\theta}_{x,k}+\lambda\mu_{k,j^{\mathrm{s}}}\hat{\theta}_{x,k}, & \text{if } \hat{y}=j^{\mathrm{t}}=j^{\mathrm{s}}\in\Lambda^{\mathrm{ts}} \end{cases} \tag{11}$$

其中 $\hat{\theta}_{x,k}$ 代表测试文档 x 属于 D^{t} 和 D^{s} 共享隐含主题 k 的概率，$\mu_{k,j^{\mathrm{s}}}$ 代表共享主题 k 与 D^{s} 的类别 j^{s} 的语义映射，$\hat{\theta}_{x,k}$ 和 $\mu_{k,j^{\mathrm{s}}}$ 揭示的即是基于共享主题的知识，从 D^{s} 迁移而来辅助对目标测试文档的主题推断和分类预测，迁移知识的作用程度由 $\lambda\in[0,1]$ 平衡因子决定.

因为迁移学习的目的是从一个或多个源任务中提取知识，并将知识应用于目标任务. 在迁移学习中，源任务和目标任务不是同等重要，迁移学习更关注目标任务，而不是同时学习源任务和目标任务[20]. 对于迁移学习任务测试文档与目标域数据集满足同分布，测试文档的预测类别应该属于目标域类标签集，而不是源数据类标签集. 因此，当 $\hat{y}=j^{\mathrm{s}}$，$j^{\mathrm{s}}\in\Lambda^{\mathrm{s}}$ 时，我们将其分类到与预测类别的主题分布最为相似的目标域类别，即 $\hat{y}=\arg\max_{j^{\mathrm{t}}}\{\mathrm{sim}(j^{\mathrm{s}},j^{\mathrm{t}})\}$，其中 $\hat{y}$ 为文档 x 的预测类别，$\mathrm{sim}(\cdot)$ 为相似度计算函数，相似性采用余弦夹角方法计算.

4 实验分析

4.1 数据集和预处理

本文使用 20newsgroup 数据集构造符合本文方法应用场景的迁移学习数据集. 20newsgroup 数据集描述如表 1 所示，共分为 20 个类别，每个类别包含 1000 篇新闻文本，其中有些相近的类别又可以组成小的数据子集，如 talk，rec 等. 本文实验数据集描述如表 2 所示，选取 20newsgroup 中 rec、sci 和 talk 三个数据子集.

迁移学习实验数据集描述如表 2 所示，每组实验数据从两个子集中选取 7 个类别作为实验数据集，其中目标域包含 5 个类别，源域包含 5 个类别，源域与目标域有 3 个交叉类别，测试集类别与目标域类别相同，每个类别随机选取 100 篇共 500 篇文档作为测试数据集. 三个数据子集两两组合共产生 rec-sci、rec-talk 和 sci-talk 三组实验数据集. 数据预处理使用 nltk. stem 词干提取，采用 TF-IDF 特征选择，保留 60% 的特征词.

表 1 20newsgroup 数据集类别描述

talk. {politics. misc, politics. guns, politics. mideast, religion. misc}
rec. {autos, motorcycles, sport. baseball, sport. hockey}
comp. {graphics, os. ms-windows. misc, sys. ibm. pc. hardware, sys. mac. hardware, windows. x}
sci. {crypt, electronics, med, space}
alt. atheism, misc. forsale, soc. religion. christian

表 2 迁移学习实验数据集类别描述

Sub-dataset	Domains	Categories
rec-sci	target domain	rec. {motorcycles, sport. baseball, sport. hockey}, sci. {crypt, space}
	source domain	rec. {autos, motorcycles, sport. baseball}, sci. {med, space}
rec-talk	target domain	rec. {motorcycles, sport. baseball, sport. hockey}, talk. {politics. guns, politics. mideast}
	source domain	rec. {autos, motorcycles, sport. baseball}, talk. {politics. misc, politics. guns}
sci-talk	target domain	sci. {crypt, electronics, space}, talk. {politics. guns, politics. mideast}
	source domain	sci. { electronics, med, space}, talk. {politics. misc, politics. guns}

4.2 实验结果及分析

实验中的对比方法简介：

(1) Tr-SLDA-TC：本文提出的基于 Tr-SLDA 迁移主题模型的 Tr-SLDA-TC 分类方法；

(2) Tr-SLDA_SVM：训练生成 Tr-SLDA 迁移主题模型，以 Tr-SLDA 主题分布表示文档向量，再用 SVM 分类；

(3) Tr-SLDA_SVM_c：训练生成 Tr-SLDA 迁移主题模型，以 Tr-SLDA 类别-主题分布表示类别向量，再用 SVM 分类；

(4) Tr-LDA-SVM：使用源域与目标域数据共同训练 LDA 主题模型后，使用 LDA 的主题分布作为特征，

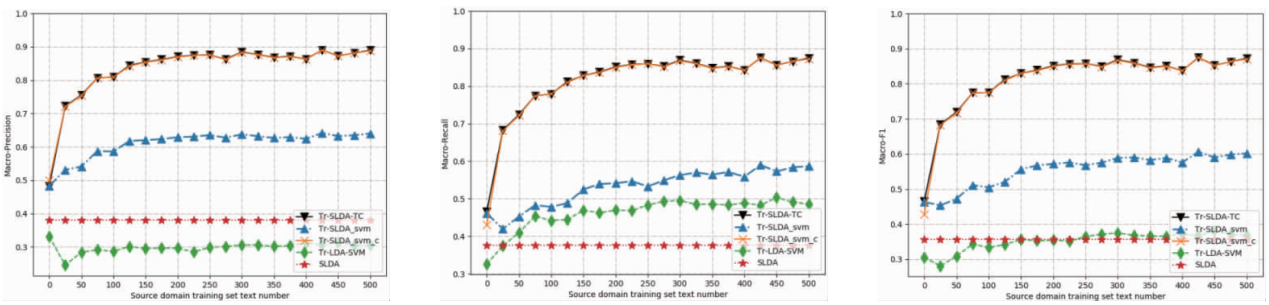
再用 SVM 分类;

(5)SLDA:为文献[7]的方法,因为此方法是有监督学习方法,所以在训练主题模型时仅使用目标域训练集而没有使用源域数据集,以此为基线对比方法.

实验结果采用宏平均精确率(Macro-Precision)、宏平均召回率(Macro-Recall)、宏平均 F1(Macro-F1)和准确率(Accuracy)四种评价指标. 模型超参数 α、β、γ 和 η 均设置为 0.01,平衡因子 λ 为 0.7,由于 Tr-SLDA-Gibbs 采样方法避免了无用主题的产生,所以主题数设定为类别总数 7 即可,迭代次数 10. 实验中每组参数进行 10 次实验取平均值作为最终结果.

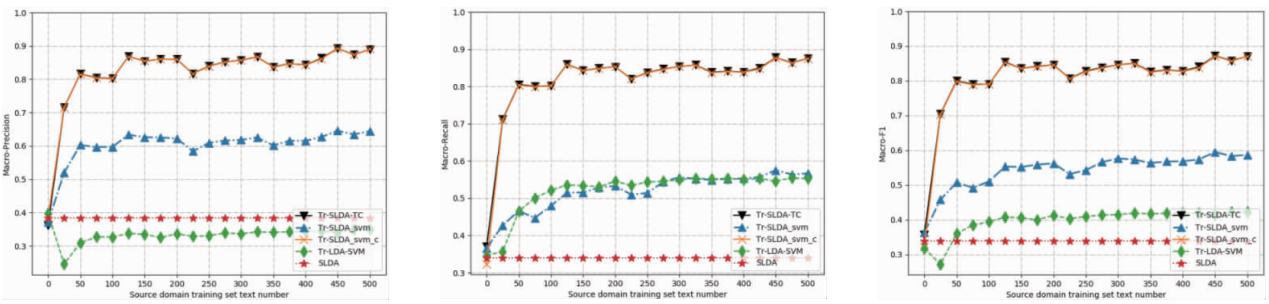
4.2.1　Tr-SLDA 的分类结果比较

在 rec-sci、rec-talk 和 sci-talk 数据子集上,固定目标域训练数据,逐渐增加源域数据,对比分析 Tr-SLDA 模型与其它分类方法的 Macro-Precision、Macro-Recall 和 Macro-F1 指标变化,实验结果如图 3 ~ 图 5 所示.



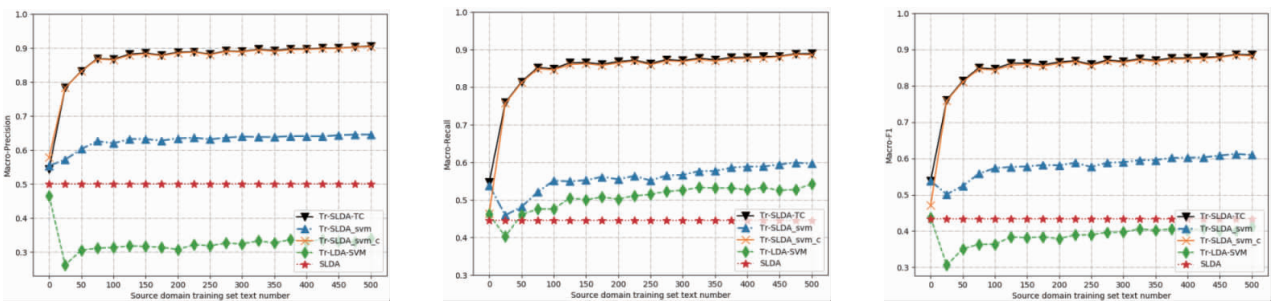
(a) Macro-Precision　　(b) Macro-Recall　　(c) Macro-F1

图3　T_n=5时各分类器分类结果随源域数据变化对比(sci-talk)



(a) Macro-Precision　　(b) Macro-Recall　　(c) Macro-F1

图4　T_n=5时各分类器分类结果随源域数据变化对比(rec-talk)



(a) Macro-Precision　　(b) Macro-Recall　　(c) Macro-F1

图5　T_n=5时各分类器分类结果随源域数据变化对比(rec-sci)

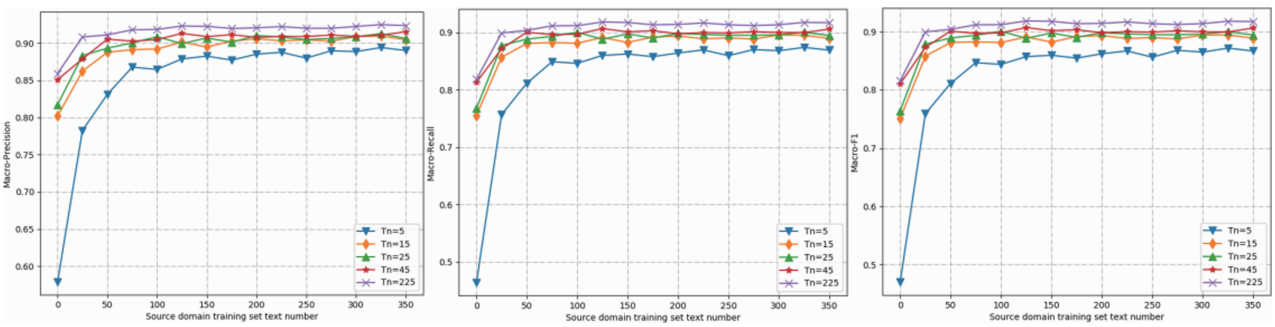
其中,横坐标表示源域训练文档数 S_n,纵坐标表示各分类指标. 固定目标域训练文档数 T_n = 5 时,随着源域数据文档数的增加,将 Tr-SLDA 模型与其它分类方法的结果对比. 由图 3 ~ 图 5 可以看出,随源域训练数据文档数 S_n 不断增大,Tr-SLDA-TC 和 Tr-SLDA_SVM_c 的各分类指标均逐渐提高,并收敛至某个上限,两方法上限几乎一致. 如图 3 所示,在 sci-talk 数据集上当 S_n = 500 时,两者的三种分类指标分别相差 0.18%、0.2% 和 0.22%. Tr-SLDA_SVM 也有大致相同的变化趋势,但分类效果始终低于 Tr-SLDA-TC 和 Tr-SLDA_SVM_c 方法. 对于 Tr-LDA-SVM 方法则受源域数据与目标域数据不满足同分布的影响使得性能略微下降,然后随着源域训练数据文档数

的增加,分类效果逐渐回升,但仍远低于基于 Tr-SLDA 主题模型的各分类方法. 如图 5 所示,在 rec-sci 数据集上当 $S_n=0$ 时,Tr-SLDA-TC 相比于 Tr-LDA-SVM 方法三种分类指标分别提升 7.87%、8.5% 和 10.14%. 而当 $S_n=500$ 时,Tr-SLDA-TC 相比于 Tr-LDA-SVM 方法三种分类指标分别提升 56.59%、34.72% 和 47.35%.

由此可见,利用目标域少量数据和部分源域数据训练生成 Tr-SLDA 迁移主题模型的有效性. Tr-SLDA 对比 Tr-LDA 可以更有效的利用源域数据,发现潜在共享主题与不同领域的类别之间的语义关联,提升面向交叉领域的文本分类性能.

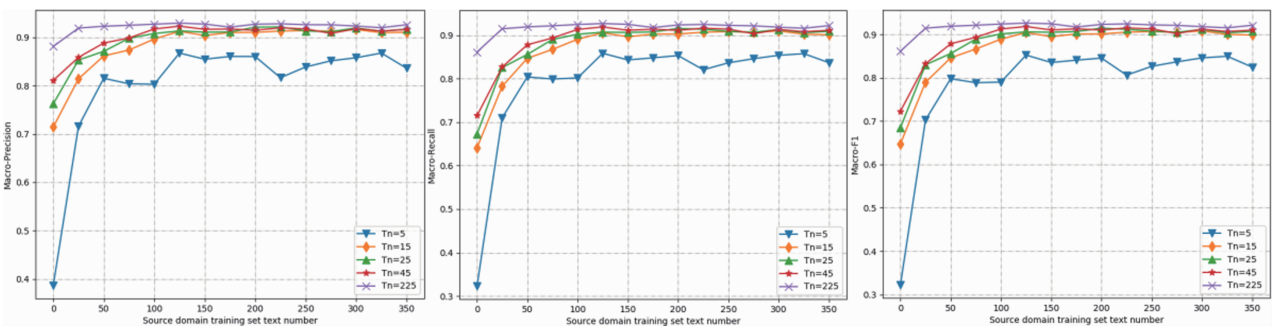
4.2.2 Tr-SLDA 迁移效果实验分析

在 rec-sci、rec-talk 和 sci-talk 数据子集上,固定不同规模的目标数据,逐渐增加源数据,对比分析 Tr-SLDA 模型的 Macro-Precision、Macro-Recall 和 Macro-F1 指标,实验结果如图 6 ~ 图 8 所示.



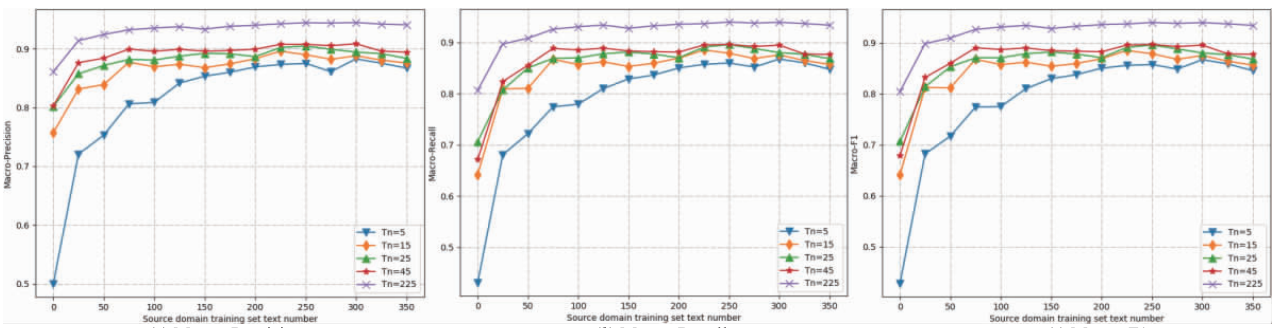
(*a*) Macro-Precision (*b*) Macro-Recall (*c*) Macro-F1

图6 不同目标数据和源数据下Tr-SLDA分类结果比较(rec-sci)



(*a*) Macro-Precision (*b*) Macro-Recall (*c*) Macro-F1

图7 不同目标数据和源数据下Tr-SLDA分类结果比较(rec-talk)



(*a*) Macro-Precision (*b*) Macro-Recall (*c*) Macro-F1

图8 不同目标数据和源数据下Tr-SLDA分类结果比较(sci-talk)

如图 6 ~ 图 8 所示,横坐标表示源域训练文档数 S_n,纵坐标分别表示 Macro-Precision、Macro-Recall 和 Macro-F1 指标. 图例中 T_n 表示目标域的训练文档数,分别固定 T_n 取 5、15、25、45 和 225,随着源域训练文档数的增加,Tr-SLDA 模型的分类结果的变化. 从图 6 ~ 图 8 可以看出:

(1)固定 T_n 后,随着源域训练数据 S_n 的增加,Tr-SLDA 模型的 Macro-Precision、Macro-Recall 和 Macro-F1 指标都逐渐提高,并收敛至某个上限,并且该上限与 T_n 成正比. 例如,从图 7 可以看出,在 rec-talk 数据子集上 $T_n=5$, $S_n=350$ 时,Tr-SLDA 模型的 Macro-Precision、Macro-Recall 和 Macro-F1 分别为 83.59%、83.66% 和 82.47%;而当 $T_n=225$, $S_n=350$ 时,Tr-SLDA 模型的

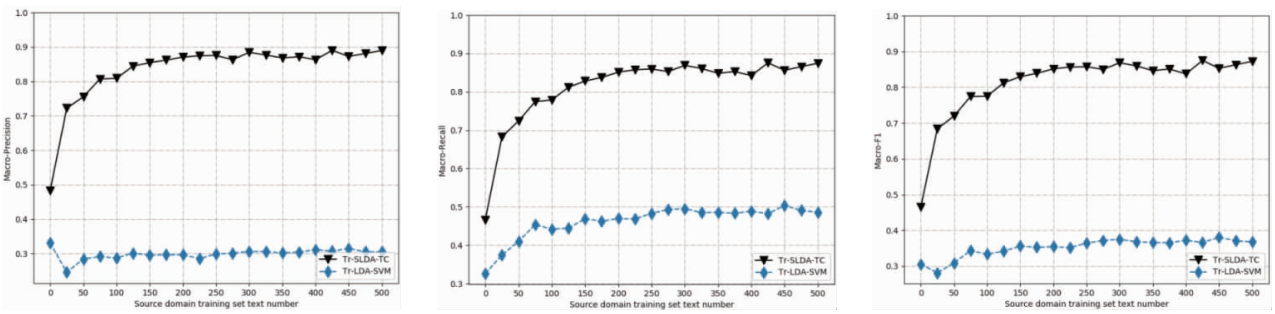
Macro-Precision、Macro-Recall 和 Macro-F1 分别为 92.6%、92.18% 和 92.12%. 由此可见,T_n 越大,Tr-SLDA 模型分类性能上限越大,即与 T_n 成正比.

(2)比较 $S_n=50$ 相对于 $S_n=0$ 时的 Tr-SLDA 模型的分类指标的提升速度,T_n 越小,Tr-SLDA 模型的分类结果各项指标的提升速度越明显,即,来自辅助源域训练数据的知识迁移效果越明显. 如图 6 所示,在 rec-sci 数据子集上,$S_n=50$ 相对于 $S_n=0$,$T_n=5$ 时,Tr-SLDA 模型的 Macro-Precision 提升 25.24%,Macro-Recall 提升 34.76%,Macro-F1 提升 34.06%,源域的辅助知识迁移效果非常明显;而 $T_n=45$ 时,Tr-SLDA 模型的 Macro-Precision 提升 5.45%,Macro-Recall 提升 8.73%,Macro-F1 提升 9.03%. 又如图 8 所示,在 sci-talk 数据子集上,$T_n=5$,$S_n=350$,Tr-SLDA 模型 Macro-Precision、Macro-Recall 和 Macro-F1 分别为 86.73%、84.72% 和 84.52%,相比 $S_n=0$ 不使用源域数据时,Tr-SLDA 模型的三种分类指标分别提升了 36.79%、41.64% 和 41.71%. 同样在 sci-talk 数据子集上,当 $T_n=225$,$S_n=350$ 时,Tr-SLDA 模型的三种分类指标分别为 94.1%、93.34% 和 93.38%,相比 $S_n=0$ 不使用源域数据时,Tr-SLDA 模型的三种指标分别提升了 8%、12.7% 和 12.98%. 由此可见,T_n 越小,Tr-SLDA 模型基于共享潜在主题,从辅助源数据的知识迁移效果越明显,即 T_n 成反比.

目标域训练数据越少,Tr-SLDA 模型从辅助源数据的知识迁移效果越明显,但目标域训练数据越多,模型随源域数据的增加最终稳定后分类性能越好. 也就是说目标域数据集的大小决定了迁移主题模型的分类性能上限,因为目标域数据集携带与测试集相同的主题信息,当目标数据训练文档数 T_n 越小,仅使用目标数据不足以训练一个好的分类器时,Tr-SLDA 迁移主题模型基于共享潜在主题,从辅助源数据的知识迁移效果越明显,从而可以验证 Tr-SLDA 迁移主题模型的有效性.

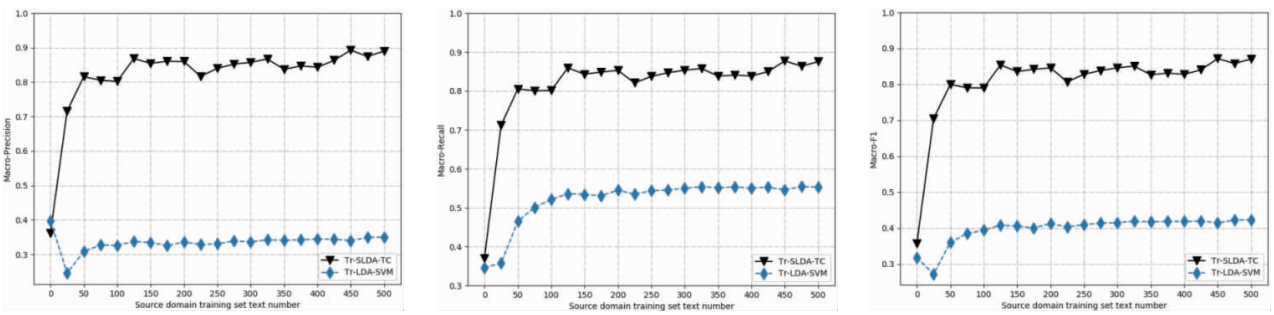
4.2.3　Tr-SLDA 与 Tr-LDA 的迁移能力比较

为了验证 Tr-SLDA 迁移主题模型和 Tr-LDA 主题模型面向交叉领域的迁移能力,在 rec-sci、rec-talk 和 sci-talk 数据子集上,固定目标域训练数据 $T_n=5$,逐渐增加源数据,对比分析 Tr-SLDA 与 Tr-LDA 的 Macro-Precision、Macro-Recall 和 Macro-F1 变化,实验结果如图 9 ~ 图 11 所示.



(a) Macro-Precision　(b) Macro-Recall　(c) Macro-F1

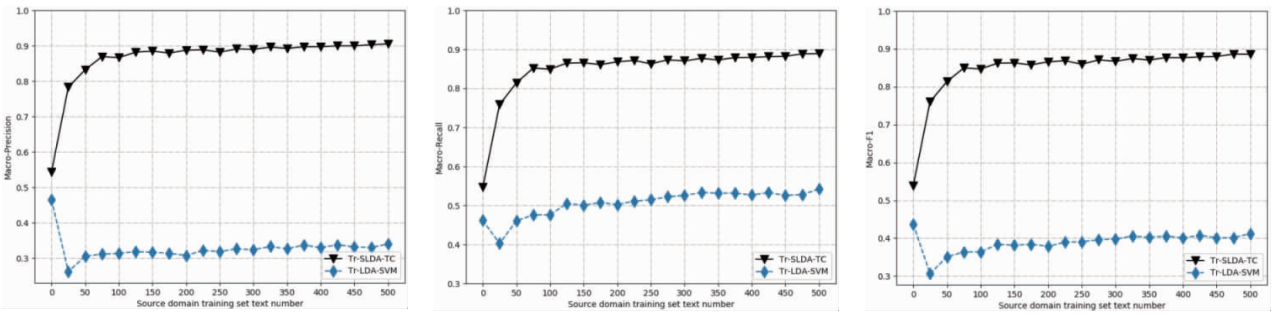
图9　T=5时Tr-SLDA与TLDA分类结果随源域数据变化对比(sci-talk)



(a) Macro-Precision　(b) Macro-Recall　(c) Macro-F1

图10　T=5时Tr-SLDA与TLDA分类结果随源域数据变化对比(rec-talk)

由图 9 ~ 图 11 可以看出,$S_n=0$,即不使用源域数据时,Tr-SLDA-TC 各分类性能指标均高于 Tr-LDA-SVM,但两者相差较小. 然而,随着源域训练数据文档数的增加,Tr-LDA-SVM 由于源域与目标域数据不满足同分布而产生主题偏移导致模型的“负迁移”现象,使得模型的整体分类性能略有下降. 如图 11 所示,在数据集 rec-sci 中,比较 $S_n=25$ 相对于 $S_n=0$ 时,Tr-LDA-SVM 三种分类指标分别下降了 25.5%、1%、14.8%,相反,Tr-SLDA-TC 的三种分类指标分别提升了 23.9%、21.2%、22.2%. 随着源域训练数据文档数的增加,Tr-LDA-SVM 各分类指标虽然有所上升,但对比 $S_n=0$,其性能并未有明显提升. 如图 10 所示,在数据集 sci-talk 上,比较 $S_n=500$ 相对于 $S_n=0$



(*a*) Macro-Precision (*b*) Macro-Recall (*c*) Macro-F1

图11 *T*=5时Tr-SLDA与TLDA分类结果随源域数据变化对比(rec-sci)

时,Macro-Precision 反倒下降 5%,Macro-Recall 和 Macro-F1 提升了 16%、6.3%,然而,Tr-SLDA-TC 的 Macro-Precision、Macro-Recall 和 Macro-F1 提升达到了 40.6%、40.8%、40.7%. 从图 9 和图 11 可以看到类似的对比.

实验表明,Tr-SLDA 迁移主题模型利用类别标签的约束,能够利用潜在共享主题从源域迁移知识,从而能够避免“负迁移”现象,而且随着源域训练文档数的增加,Tr-SLDA-TC 的分类性能明显提升. 而 Tr-LDA 迁移主题模型是无监督学习,没有有效利用源域与目标域的类别与潜在主题之间的知识关联,从而产生主题偏移导致“负迁移”. 因此,面向交叉领域,Tr-SLDA 迁移主题模型的迁移能力更强.

通过大量的实验对比分析,Tr-SLDA 迁移主题模型可有效利用源域数据,发现潜在共享主题与不同领域类别之间的语义关联,提升面向交叉领域的文本分类性能,有效避免了“负迁移”现象. Tr-SLDA-TC 在源域数据的辅助下分类性能远高于基线对比方法 SLDA 和 Tr-LDA-SVM 分类方法.

5 结论

融合 SLDA 和迁移学习提出了一种迁移主题模型 Tr-SLDA,利用标签的约束将源域与目标域数据映射到一个潜在的共享主题空间,提出新的 Tr-SLDA-Gibbs 主题采样和参数估计方法,旨在通过发现潜在的共享主题与不同领域的类别之间的语义关联,将源域的知识迁移到目标域分类任务中,从而能够有效解决“负迁移”问题. 进而基于迁移主题模型,提出 Tr-SLDA-TC 文本分类算法. 实验表明,当目标领域数据不足时,Tr-SLDA-TC 能够利用源域的迁移知识提高目标域分类任务的精度. 下一步,我们将结合迁移学习方法对多标签文本数据下的主题模型及分类方法展开研究.

参考文献

[1] Dai W,Jin O,Xue G R,et al. Eigentransfer:a unified framework for transfer learning[A]. Proceedings of the 26th Annual International Conference on Machine Learning[C]. New York, NY: Association for Computing Machinery, 2009. 193 – 200.

[2] Pan W,Yang Q. Transfer learning in heterogeneous collaborative filtering domains[J]. Artificial Intelligence,2013,197(4):39 – 55.

[3] Long M,Wang J,Ding G,Shen D,Yang Q. Transfer learning with graph co-regularization[J]. IEEE Transactions on Knowledge & Data Engineering,2014,26(7):1805 – 1818.

[4] Zhang X,Yang Q. Transfer hierarchical attention network for generative dialog system[J]. International Journal of Automation and Computing,2019,16(6):720 – 736.

[5] Pan W,Yang Q,Duan Y,et al. Transfer learning for semi-supervised collaborative recommendation[J]. The ACM Transactions on Interactive Intelligent Systems, 2016, 6(2):1 – 21.

[6] 吴斌,吉佳,孟琳,等. 基于迁移学习的唐诗宋词情感分析[J]. 电子学报,2016,44(11):2780 – 2787.
Wu B,Ji J,Meng L,et al. Transfer learning based sentiment analysis for poetry of the tang dynasty and song dynasty[J]. Acta Electronica Sinica,2016,44(11):2780 – 2787. (in Chinese)

[7] Zhao G L,Xiang Y L,Li M Q,et al. A cross-region transfer learning method for classification of community service cases with small datasets[J]. Knowledge-Based Systems, 2020,193:105390.

[8] Blei D M,Ng A,Jordan M. Latent dirichlet allocation[J]. Journal of Machine Learning Research,2003,3(4 – 5):993 – 1022.

[9] Wood J,Tan P,Wang W,et al. Source-LDA:enhancing probabilistic topic models using prior knowledge sources[A]. International Conference on Data Engineering[C]. San Diego,California:IEEE,2017. 411 – 422.

[10] Zhang H,Zhong G. Improving short text classification by learning vector representations of both words and hidden topics[J]. Knowledge-Based Systems,2016,102(6):76 – 86.

[11] Kandemir M,Kekeç T,Yeniterzi R. Supervising topic models with gaussian processes[J]. Pattern Recognition,

2018,77(5):226 – 236.

[12] 石敏,刘建勋,周栋,等. 基于多重关系主题模型的 Web 服务聚类方法[J]. 计算机学报,2019,42(04):820 – 836.
Shi M,Liu J X,Zhou D,et al. Multi-relational topic model-based approach for web services clustering[J]. Chinese Journal of Computers,2019,42(04):820 – 836. (in Chinese)

[13] 张雄,陈福才,黄瑞阳. 基于双词主题模型的半监督实体消歧方法研究[J]. 电子学报,2018,46(3):607 – 613.
Zhang X,Chen F C,Huang R Y. Semi-supervised entity disambiguation method research based on bi term topic model[J]. Acta Electronica Sinica,2018,46(3):607 – 613. (in Chinese).

[14] Mei Q,Ling X,Wondra M,et al. Topic sentiment mixture:modeling facets and opinions in weblogs[A]. Proceedings of the 16th International World Wide Web Conference (WWW′07). Association for Computing Machinery[C]. New York,NY,USA,2007. 171 – 180.

[15] Niu Z,Hua G,Wang L,et al. Knowledge based topic model for unsupervised object discovery and localization [J]. IEEE Transactions on Image Processing,2018,27(1):50 – 63.

[16] Chen C,Zare A,Trinh H N,et al. Partial membership latent dirichlet allocation for soft image segmentation[J]. IEEE Transactions on Image Processing,2017,26(12):5590 – 5602.

[17] 刘杰,杜军平. 基于潜在主题融合的跨媒体图像语义标注[J]. 电子学报,2014,42(05):987 – 991.
Liu J,Du J P. Latent topic fusion-based cross-media image semantic annotation[J]. Acta Electronica Sinica,2014,42(05):987 – 991. (in Chinese)

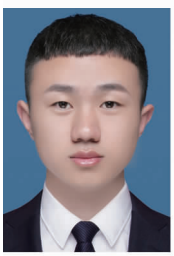
[18] 唐焕玲,窦全胜,于立萍,宋英杰,鲁明羽. 有监督主题模型的 SLDA-TC 文本分类新方法[J]. 电子学报,2019,47(06):1300 – 1308.
Tang H L,Dou Q S,et al. SLDA-TC:a novel text categorization approach based on supervised topic model[J]. Acta Electronica Sinica,2019,47(06):1300 – 1308. (in Chinese)

[19] 李永忠,郑滔. 基于标签的半监督 HDP 文本分类主题模型[J]. 模式识别与人工智能,2017,30(12):1138 – 1148.
LI Yongzhong,ZHENG Tao. Semi-supervised labeled hierarchical dirichlet process topic model for document categorization[J]. Pattern Recognition and Artificial Intelligence,2017,30(12):1138 – 1148. (in Chinese)

[20] Pan S J,Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering,2010,22(10):1345 – 1359.

作者简介

唐焕玲　女,教授,博士,硕士生导师,1970 年生于山东龙口. 2004 年于清华大学获得工学硕士学位,2009 年于大连海事大学大学获得工学博士学位. 从事机器学习、人工智能、数据挖掘等方向的理论及应用研究.
E-mail:thl01@163.com



郑　涵　男,1997 年生于河南商丘. 山东工商学院硕士研究生,主要研究方向为机器学习、人工智能、数据挖掘.
E-mail:zhenghan0503@163.com

刘艳红　女,1995 年生于山东烟台. 山东工商学院硕士研究生,研究方向为机器学习与数据挖掘.
E-mail:2669349709@qq.com

马思源　女,1994 年生于河南周口. 山东工商学院硕士研究生,主要研究方向为计算机视觉.
E-mail:masiyuan423@163.com

窦全胜　男,教授,博士,硕士生导师. 1971 年生于黑龙江大庆. 2001 年、2005 年于吉林大学分别获得理学硕士学位、工学博士学位. 从事人工智能、机器学习、演化计算等方向的理论及应用研究.
E-mail:li_dou@163.com

鲁明羽　男,教授,博士生导师. 1963 年生于黑龙江鸡西. 1988 年、2002 年于清华大学分别获得工学硕士和工学博士学位,从事机器学习、人工智能、数据挖掘等方向的理论及应用研究.
E-mail:lumingyu@dlmu.edu.cn