

融合主题模型和动态路由的小样本学习方法

张淑芳¹, 唐焕玲^{1,3,4}, 郑涵², 刘孝炎¹, 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, 烟台 264005; 2. 山东工商学院信息与电子工程学院, 烟台 264005; 3. 山东省高等学校协同创新中心: 未来智能计算, 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), 烟台 264005; 5. 大连海事大学信息科学技术学院, 大连 116026)

摘要: 针对小样本学习标注训练样本过少, 导致特征表达力弱的问题, 本文结合有监督主题模型(Supervised LDA, SLDA)和动态路由算法提出一种新的动态路由原型网络模型(Dynamic routing prototypical network based on SLDA, DRP-SLDA)。利用SLDA主题模型建立词汇与类别之间的语义映射, 增强词的类别分布特征, 从词粒度角度编码获得样本的语义表示。提出动态路由原型网络(Dynamic routing prototypical network, DR-Proto), 通过提取交叉特征利用样本之间的语义关系, 采用动态路由算法迭代生成具有类别代表性的动态原型, 旨在解决特征表达问题。实验结果表明, DRP-SLDA模型能有效提取词的类别分布特征, 且获取动态原型提高类别辨识力, 从而能够有效提升小样本文本分类的泛化性能。

关键词: 小样本学习; 元学习; 原型网络; 有监督主题模型; 文本分类

中图分类号: TP181 **文献标志码:** A

Few-Shot Learning Method Based on Topic Model and Dynamic Routing Algorithm

ZHANG Shufang¹, TANG Huanling^{1,3,4}, ZHENG Han², LIU Xiaoyan¹, DOU Quansheng^{1,3,4}, LU Mingyu⁵

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

Abstract: Aiming at the problem that the training samples for few-shot learning are too few, which leads to the weak expression of features, a novel dynamic routing prototypical network based on SLDA (DRP-SLDA) is proposed based on the supervised topic model (Supervised LDA, SLDA) and dynamic routing algorithm. The SLDA topic model is used to establish the semantic mapping between words and categories, enhance the category distribution characteristics of words, and obtain the semantic representation of samples from the perspective of word granularity. The dynamic routing prototypical network (DR-Proto) is presented. The network makes full use of the semantic relationship between samples by extracting cross features, and uses the dynamic routing algorithm to iteratively

generate dynamic prototype with category representation, so as to solve the problem of feature expression. The experimental results show that the DRP-SLDA model can effectively extract the category distribution characteristics of words and dynamically obtain the dynamic prototype to increase the category identification, which can obviously improve the generalization ability of few-shot text classification.

Key words: few-shot learning; meta-learning; prototypical networks; supervised LDA; text categorization

引 言

有监督深度学习^[1-3]需要大量的标注样本,然而标注样本获取代价昂贵;相反,人类能够在样本匮乏的情况下学习新知识。因此小样本学习^[4-5](Few-shot learning, FSL)方法的研究得到广泛关注,其目标是在每个类别只有少量样本的情况下训练模型,但是在无法有效提取特征时,导致模型泛化性能下降,是小样本学习亟待解决的问题。早期研究者应用迁移学习^[6]微调预训练模型。近年来,元学习^[7]能够比较好地解决小样本学习问题,主要有基于度量的算法、基于数据增强的算法和基于模型优化的算法^[8]。其中,基于度量的小样本学习简单有效,在面向图像和文本的任务中得到广泛关注。Vinyals等^[9]提出匹配网络,通过分段训练策略迁移元知识^[10]生成以余弦距离度量的加权K-近邻分类器。Snell等^[11]提出原型网络,假设每个类的样本在度量空间中都存在原型,同属一个类别的样本均值向量作为该类的原型。Fort等^[12]提出高斯原型网络,将样本表示为高斯协方差矩阵,利用样本权重构造与类相关的度量函数。Gao等^[13]提出基于注意力的混合原型网络,设计了样本级别和特征级别的两种注意力机制,分别捕捉对分类更重要的样本和特征,提高关系分类模型在噪声数据集上的性能和鲁棒性。Sun等^[14]提出层次注意力原型网络,构建了特征、单词和样本3种级别的注意力机制,分别赋予不同的权重分数,训练每个类的原型。度量学习应用于自然语言处理的不同领域,如单词预测、知识图谱以及任务型对话系统。有效提取不同粒度的词特征和样本特征,是基于度量的小样本学习需要解决的难点问题。本文从词粒度考虑不同类别的词分布特征,动态更新每类的原型,提出一种基于SLDA^[15]的动态路由原型网络模型(Dynamic routing prototypical network based on SLDA, DRP-SLDA)。利用SLDA主题模型识别词与类别之间的精准语义映射,获得每个类别的词分布特征。基于支持集和查询集,提出动态路由原型网络(Dynamic routing prototypical network, DR-Proto),旨在通过动态路由算法^[16]有效利用样本之间语义交叉特征训练动态原型,从而提升小样本文本分类的泛化性能。

1 DRP-SLDA 模型

1.1 DRP-SLDA 模型框架

C-way K-shot表示含有C个类别且每类有K个标记样本的小样本分类问题。令 $D = \{D_{\text{train}}, D_{\text{test}}\}$ 且 $D_{\text{train}} \cap D_{\text{test}} = \emptyset$ 。从训练集 D_{train} 选C个类别,每类选K个有标签样本 $S = \{(x_i, y_i)\}_{i=1}^m$ ($m = C \times K$),作为支持集, x_i 表示第*i*个样本, $y_i \in [1, C]$ 为 x_i 的类标签;每类另外选L个无标签样本 $Q = \{(x_q, y_q)\}_{q=1}^h$ ($h = C \times L$)作为查询集, x_q 表示第*q*个样本, $y_q \in [1, C]$ 为 x_q 的类标签。测试集 D_{test} 由C类的有标签支持集和无标签查询集构成。元学习方法仅从 D_{train} 中采样1组支持集和查询集,样本过少模型难以学习良好特征表示。借鉴文献[17]引入源集辅助训练,增强特征表示,如图1所示。以20newsgroup数据集的3-way 2-shot任务为例,选3个类作为 D_{test} ,其余17个类作为 D_{train} 。元训练时,从 D_{train} 选3个类采样支持集和查询集,剩余14个类作为源集。元测试时,从 D_{test} 中采样支持集和查询集,

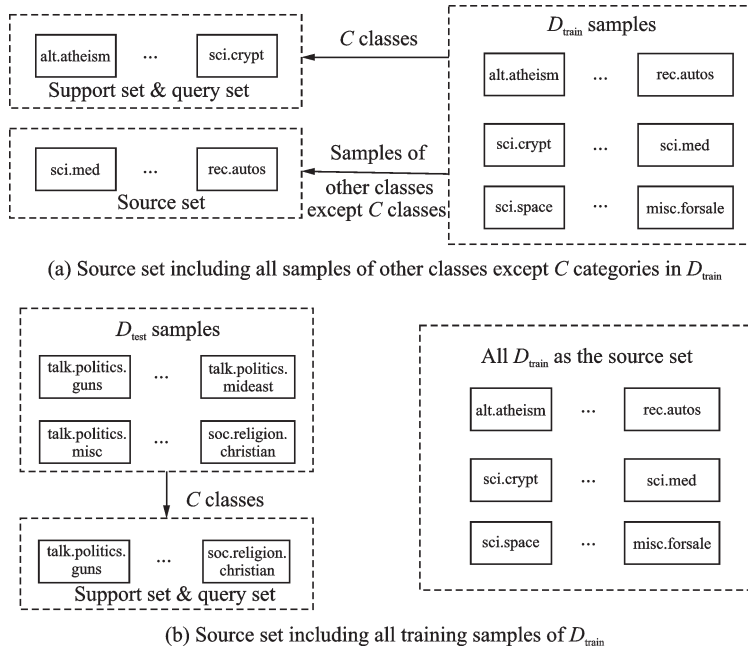


图1 元训练样本

Fig.1 Meta-training samples

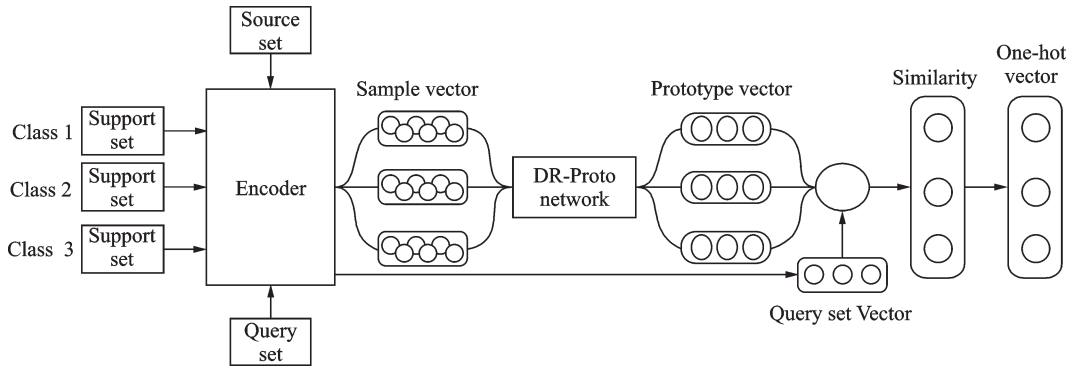


图2 DRP-SLDA模型的 C -way K -shot($C=3, K=2$)框架

Fig.2 C -way K -shot ($C=3, K=2$) architecture of DRP-SLDA model

将 D_{train} 中所有类的样本作为源集。DRP-SLDA 模型框架如图 2 所示,由 Encoder 和 DR-Proto network 构成。其中,Encoder 利用 SLDA 主题模型增强词的分布特征,得到词的特征编码。DR-Proto network 为本文提出的动态路由原型网络,利用动态路由算法提取支持集和查询集的样本交叉特征计算每类的原型。通过欧氏距离 $d(\cdot)$ 度量查询集与原型的相似性以分类预测。

1.2 基于 SLDA 编码方法

为获得每个类别的词分布特征,本文提出基于 SLDA 主题模型的词编码方法,即图 2 所示 DRP-SLDA 模型框架中 Encoder。

1.2.1 SLDA 主题模型

SLDA 主题模型的概率图模型如图 3 所示。支持集 $S = \{(x_i, y_i)\}_{i=1}^m$, x_i 表示第 i 个样本, y_i 为 x_i 的

标签。其中 m 为样本数, C 为类别数, T 为主题数, l 为第 i 个样本的长度。 x_i^w 表示第 i 个样本中第 w 个词, z_i^w 表示分配给 x_i^w 的主题, $\mathbf{z} = (z_1, \dots, z_V)$ 为主题向量 (V 为单词数)。 θ_i 为样本 i 的主题概率分布, φ_t 为第 t 个主题的词概率分布, δ_t 为第 t 个主题类别概率分布。 θ_i, φ_t 和 δ_t 服从 Dirichlet 分布, 其先验参数为 α, β, γ 。 获得词 x_i^w 的主题标号 z 后, 假设 $x_i^w = w$ 且所在样本 i 的类标签 $y_i = j, j \in [1, C]$, 通过统计频次计算 θ_i, φ_t 和 δ_t , 有

$$\theta_{i,t} = (n_i^{(t)} + \alpha_t) / \sum_{z=1}^T (n_i^{(z)} + \alpha_z) \quad (1)$$

$$\varphi_{t,w} = (n_t^{(w)} + \beta_w) / \sum_{v=1}^V (n_t^{(v)} + \beta_v) \quad (2)$$

$$\delta_{t,j} = (n_t^{(j)} + \gamma_j) / \sum_{c=1}^C (n_t^{(c)} + \gamma_c) \quad (3)$$

式中: $\theta_{i,t}$ 表示样本 i 的词分配给主题 t 的概率; $\varphi_{t,w}$ 表示主题 t 分配给词 w 的概率; $\delta_{t,j}$ 表示主题 t 属于类别 j 的概率。 $n_i^{(t)}$ 表示第 i 个样本被赋予主题 t 的次数; $n_t^{(w)}$ 表示词 w 被赋予主题 t 的次数; $n_t^{(j)}$ 表示类别 j 被赋予主题 t 的次数。 α_t 为主题 t 的 Dirichlet 先验; β_w 为词 w 的 Dirichlet 先验; γ_j 为类别 j 的 Dirichlet 先验。 $t=1, \dots, T; w=1, \dots, V; j=1, \dots, C$ 。

1.2.2 基于SLDA的编码表示

小样本学习因样本匮乏难以较好学习词特征, 由于SLDA主题模型的 φ 和 δ 识别词汇与类别之间的精准映射, 基于SLDA的Encoder旨在从单词-类别分布中增强特征表示学习, 如图4所示。SLDA模型提取单词在特定类分布特征, 源集获取单词的通用性特征表示。通过双向长短期记忆网络 (Bi-directional long short-term memory, BiLSTM)^[18] 计算单词权重, 将 word2vec 后的单词加权求和获得样本特征表示。

基于SLDA计算单词对于特定类的权重为

$$S = ((-\delta^T \varphi) \log(\delta^T \varphi))^{-1T} \quad (4)$$

式中: $S \in \{[0, 1]\}^{l \times C}$, l 表示每个样本中含有的单词总数, C 表示类别总数; $\delta \in \mathbf{R}^{T \times C}$ 表示主题的类别概率分布; $\varphi \in \mathbf{R}^{T \times l}$ 表示主题的词概率分布, T 是主题数。由 $\delta^T \varphi$ 计算类别-一词的精准映射, 从而获得单词特定类中的分布特征。信息熵衡量词分类的不确定性, 熵值 $[0, 1]$ 越大越难判别所属类别。 $S(x_i^w)$ 表示第 i 个样本的第 w 个单词是第 j 类的概率值, $S(x_i^w)$ 越大, 词 x_i^w 在特定类中重要性越大。

提取源集中单词通用性特征^[19]有

$$R(x_i^w) = \frac{\mu}{P(x_i^w) + \mu} \quad (5)$$

式中: x_i^w 表示第 i 个样本中的第 w 个词; μ 为平滑系数; $P(\cdot)$ 表示词频。 $R(x_i^w)$ 降低源集高频词的权重, 获得 x_i^w 在通用类中的重要性。则 R 表示为

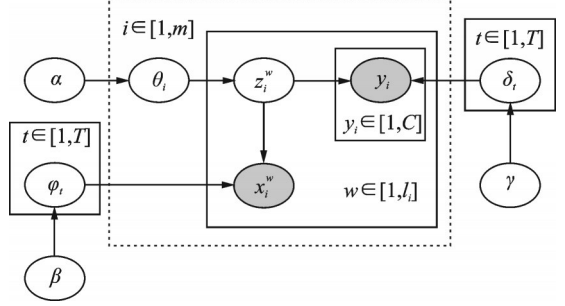


图3 SLDA 概率图模型

Fig.3 SLDA probabilistic graphical model

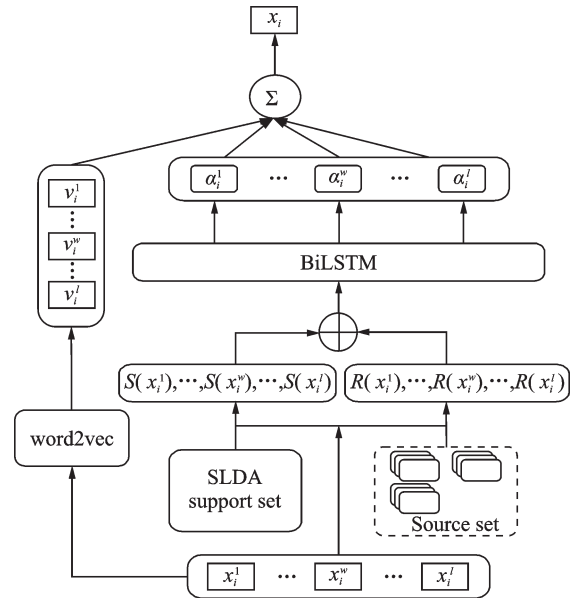


图4 基于SLDA的Encoder架构

Fig.4 Encoder architecture based on SLDA

$$R = \begin{Bmatrix} \{R(x_i^1)\}^C \\ \{R(x_i^2)\}^C \\ \vdots \\ \{R(x_i^l)\}^C \end{Bmatrix}^{l \times C}$$

S 和 R 互补难以设计合理的组合函数,且两者对分类重要性一致。关键词与词序和语境有关,通过 BiLSTM 融合拼接 S 和 R ,有

$$H_i = \text{BiLSTM}(\text{Concat}(R, S)) \quad (6)$$

式中: H_i 表示样本 i 全部单词的特征表示, $H_i \in \mathbf{R}^{l \times d}$ 。

令 $H_i = \{h_i^1, h_i^2, \dots, h_i^l\}$ 为样本 i 全部单词的特征表示, $H_i \in \mathbf{R}^{l \times d}$, 则第 i 个样本的单词权重为

$$\alpha_i = \text{softmax}(\tanh(\mathbf{w} \cdot H_i^T + \mathbf{b})) \quad (7)$$

式中: $\mathbf{w} \in \mathbf{R}^{1 \times d}$ 、 $\mathbf{b} \in \mathbf{R}^{1 \times l}$ 为可学习参数; $\tanh(\cdot)$ 为激活函数。通过 $\text{softmax}(\cdot)$ 归一化获得第 i 个样本 l 个单词的权重 $\alpha_i \in \mathbf{R}^{1 \times l}$ 。

为单词赋予不同权重,则有

$$x_i^w = \sum_{u=1}^d \alpha_i^{w,u} \cdot v_i^w \quad (8)$$

式中: $v_i^w \in \mathbf{R}^{1 \times d}$ 表示词 x_i^w 经过 word2vec 的低维向量; $\alpha_i^{w,u}$ 为权重,通过加权求和获得样本标量值 x_i^w , 令 $\mathbf{x}_i = \{x_i^1, \dots, x_i^w, \dots, x_i^l\}$, 则样本向量 $\mathbf{x}_i \in \mathbf{R}^{l \times 1}$ 。

Encoder 利用 SLDA 从词粒度角度引入主题特征编码,获得语义增强的样本表示。

1.3 DR-Proto 网络

基于 SLDA 的 Encoder, 提出动态路由原型网络 DR-Proto。利用支持集和查询集的样本语义交叉特征获得原型,使分类边界更清晰。DR-Proto 网络如图 5 所示,即图 2 所示 DRP-SLDA 模型框架中的 DR-Proto network。图 5 中,DR-Proto 网络提取支持集和查询集的样本交叉特征 ψ_i^j , 通过动态路由算法调整耦合系数 $m_g^{i,j}$, 更新样本权重 $\eta_g^{i,j}$, 多次迭代获得动态原型 C^j 。通过特征共享变换^[20]将 Encoder 的 x_i^j 和 x_q 映射到同一空间,有

$$\tilde{x}_i^j = \text{Squash}(W_s \cdot x_i^j + b_s) \quad (9)$$

$$\tilde{x}_q = \text{Squash}(W_s \cdot x_q + b_s) \quad (10)$$

式中: $x_i^j \in \mathbf{R}^{l \times 1}$ 为支持集类 j 的第 i 个样本; $x_q \in \mathbf{R}^{l \times 1}$ 为查询集第 q 个样本; $W_s \in \mathbf{R}^{l \times l}$ 、 $b_s \in \mathbf{R}^{l \times 1}$ 为共享参数。通过特征映射使 \tilde{x}_i^j 、 $\tilde{x}_q \in \mathbf{R}^{l \times 1}$ 距离尽可能近, $i \in [1, K]$, $j \in [1, C]$, $q \in [1, h]$ 。Squash $(\cdot)^{[21]}$ 使向量 x 方向不变且大小取值范围 $[0, 1)$, 即

$$\text{Squash}(x) = \frac{\|x\|^2}{1 + \|x\|^2} \frac{x}{\|x\|} \quad (11)$$

考虑支持集和查询集满足独立同分布,通过提取交叉特征 ψ_i^j 以利用文本语义关系,有

$$\psi_i^j = \tilde{x}_i^j \odot \tilde{x}_q + |\tilde{x}_i^j - \tilde{x}_q| \quad (12)$$

式中: \odot 为元素积; $|\cdot|$ 表示取绝对值。若 \tilde{x}_i^j 和 \tilde{x}_q 属于同类,样本间的距离近,则交叉特征多;否则,样本间的距离远,则交叉特征少。通过交叉特征 $\psi_i^j \in \mathbf{R}^{l \times 1}$, 使样本特征相应增强或弱化。

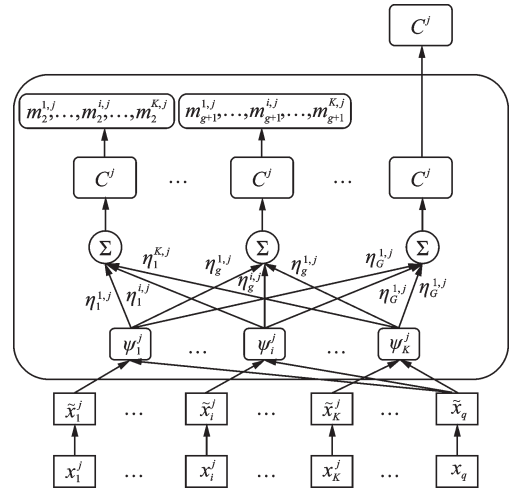


图5 DR-Proto 网络架构

Fig.5 DR-Proto network architecture

计算权重 η_g^j , 获得不同样本对原型的重要性, 即有

$$\eta_g^j = \text{softmax}(m_g^j) \quad (13)$$

式中耦合系数 $m_g^j \in \mathbf{R}^{1 \times K}$ 为先验知识, $g = 1, \dots, G$ 为动态路由迭代次数。每个样本训练之前等概率影响原型, 从而初始化 $m_g^j = 0$ 使 $\eta_g^j \in \mathbf{R}^{1 \times K}$ 均匀分布。

根据权重计算每类的原型, 有

$$C^j = \text{Squash}\left(\sum_{i=1}^K (\eta_g^{i,j} \cdot \psi_i^j)\right) \quad (14)$$

式中: $\eta_g^{i,j}$ 为第 g 次迭代时类 j 第 i 个样本的权重; K 为每个类别的样本个数。 ψ_i^j 为交叉特征计算如式 (12), 得到类 j 的原型 $C^j \in \mathbf{R}^{l \times 1}$ 。

为了增大同类样本相关性而降低不同类相关性, 自上而下调整耦合系数, 有

$$m_{g+1}^{i,j} = m_g^{i,j} + (\tilde{x}_i^j)^T \cdot C^j \quad (15)$$

式中: 若 \tilde{x}_i^j 属于原型 C^j , $(\tilde{x}_i^j)^T \cdot C^j$ 则增大耦合系数, 增强该样本对原型的影响; 若 \tilde{x}_i^j 不属于原型 C^j , $(\tilde{x}_i^j)^T \cdot C^j$ 则减小耦合系数, 使该样本的影响边缘化。

由融合动态路由算法^[16]的式(12~15)可知, 利用语义关联的支持集和查询集提取交叉特征 ψ_i^j , 进而采用权重机制获得动态原型 C^j 。DR-Proto网络通过为样本特征赋予权重获得动态原型, 从而改进均值原型^[11]有效提取类别特征。

1.4 分类预测

样本 x_q 与第 j 类原型的相似度计算采用欧氏距离, 有

$$d(x_q, C^j) = \sum_{w=1}^l (C_w^j - x_q^w)^2 \quad (16)$$

对样本 x_q 的类别预测, 有

$$p(\hat{y}_{j,q} | x_q) = \frac{\exp(-d(x_q, C^j))}{\sum_{j=1}^C \exp(-d(x_q, C^j))} \quad (17)$$

式中: $d(x_q, C^j)$ 表示样本 x_q 与第 j 类原型的欧氏距离; $\hat{y}_{j,q}$ 表示 x_q 预测为类别 j 的概率。 $w \in [1, l]$, $j \in [1, C]$ 。

采用均方误差损失 (Mean squared error, MSE) 优化参数, 有

$$\text{MSE}(y, \hat{y}) = \frac{1}{h} \sum_{j=1}^C \sum_{q=1}^h (y_q - \hat{y}_{j,q})^2 \quad (18)$$

式中: y_q 表示查询集 x_q 的真实标签 one-hot 编码值; h 表示查询集样本数。

2 实验分析

2.1 实验数据集

为验证模型有效性, 利用 20newsgroup 英文数据集、FewRel 关系数据集和 Sogou 中文数据集, 采样支持集和查询集, 构建 C -way K -shot 分类任务进行对比实验, 数据集描述如表 1 所示。

2.2 实验结果及分析

2.2.1 对比实验结果及分析

对比模型有: (1) Finetune^[22]: 有监督的线性微调分类器。(2) 1-nearest neighbor classifier: 有监督的最近邻分类器。(3) Prototypical network^[11]: 每类支持集样本的均值特征向量作为类原型的原型网络。(4) MAML (Model-agnostic meta-learning)^[23]: 通过梯度求和优化不同子任务初始化参数的元学习模

表1 数据集描述

Table 1 Description of data sets

Data set	Category of train set	Category of test set
20newsgroup	comp.{graphics, os.ms-windows.misc, windows.x sys.ibm.pc.hardware, sys.mac.hardware, }	talk.{politics.misc, politics.guns, politics.guns, religion.misc} soc.religion.christian
	rec.{autos, motorcycles, sport.baseball, sport. hockey}	
	sci.{crypt, electronics, med, space}	
	alt.atheism misc.forsale	
FewRel	P6, P22, P27, P39, P57, P84, P123, P140, P150, P175, P449, P495, P527, P706, P740	P931, P1344, P1435, P1923, P4552
Sogou	军事, 健康, 旅游, 教育, IT, 财经	体育, 文化, 招聘

型。(5) RR-D2_LR-D2^[24]:采用岭回归和逻辑回归的元学习模型。(6) Distributional signatures^[17]:将词汇特征映射成注意力分数衡量文本表示,并使用岭回归器分类预测的元学习模型。

DRP-SLDA模型与上述6种模型在FewRel和20newsgroup数据集^[17]上的分类结果比较如表2所示。

表2 模型在FewRel数据集和20newsgroup数据集上的准确率

Table 2 Accuracy of the model on the FewRel and 20newsgroup data sets

模型	FewRel		20newsgroup		%
	5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot	
Finetune	54.1	71.1	33.0	47.1	
1-nearest neighbor	46.8	60.6	38.8	51.9	
Prototypical networks	43.0	61.9	37.8	46.5	
MAML	48.2	65.8	37.2	48.6	
R2-D2_LR-D2	55.6	75.3	44.8	64.3	
Distributional signatures	67.1	83.5	52.1	68.3	
DRP-SLDA	71.7	88.7	60.7	84.9	

由表2可知,相比有监督学习的Finetune模型和1-nearest neighbor模型,元学习的模型整体效果表现良好。Prototypical networks模型使用欧式距离度量映射空间内查询集与原型之间的距离预测分类;MAML模型放弃距离度量方式,通过在多个子任务中使用梯度下降法训练初始参数,微调参数以计算不同子任务的损失快速收敛模型,其在FewRel数据集上的准确率分别是48.2%、65.8%。然而,MAML模型缺乏针对各任务的分析。为此,RR-D2_LR-D2模型对不同任务生成先验知识,以可微分回归作为基分类器。Distributional signatures模型在RR-D2_LR-D2模型的基础上,将词分布特征转成注意力分数,用岭回归分类预测,在两种数据集上,两种小样本分类任务的准确率都得以提高,如20newsgroup数据集上分别提高7.3%、4.0%。

DRP-SLDA模型对比模型(1~5)在20newsgroup数据集上准确率提升10%~30%,表明DRP-SLDA模型引入源集是有效的,能从不同角度提取词汇特征。DRP-SLDA模型相较于Distributional signatures模型在20newsgroup数据集上准确率分别提高8.6%、16.6%。原因在于Distributional signatures模型仅考虑词汇特征没有考虑样本权重对分类的影响,而DRP-SLDA模型利用SLDA模型增强词分布特征且DR-Proto网络为样本赋予权重获得动态原型。

2.2.2 DRP-SLDA 模型消融分析

消融方法简介:(1) DRP-SLDA:本文提出的小样本分类模型。(2) DRP-SLDA_NR:消融 DRP-SLDA 模型中源集在数据扩充上的影响。(3) DRP-SLDA_NS:消融 DRP-SLDA 模型中 SLDA 主题模型对词分布特征的影响。(4) DRP-SLDA_ND:消融 DRP-SLDA 模型中交叉特征 ψ 的影响。(5) DRP-SLDA_NDR:消融 DRP-SLDA 模型中源集数据扩充和交叉特征 ψ 的共同影响。(6) DRP-SLDA_NDS:消融 DRP-SLDA 模型中 SLDA 主题模型和交叉特征 ψ 的共同影响。

为了分析 DRP-SLDA 模型不同组件对分类效果的影响,在 Sogou 和 20newsgroup 数据集上分别进行 3-way 1-shot,5-way 5-shot 消融实验,实验结果如表 3 和图 6,7 所示。其中,Macro-precision 为宏精确度;Macro-recall 为宏召回率;Macro-F₁ 为 $2 \times \text{Macro-precision} \times \text{Macro-recall} / \text{Macro-precision} + \text{Macro-recall}$ 。

表 3 消融方法在 Sogou 数据集和 20newsgroup 数据集上的实验结果

方法	3-way 1-shot(Sogou)			5-way 5-shot(20newsgroup)		
	Macro-precision	Macro-recall	Macro-F ₁	Macro-precision	Macro-recall	Macro-F ₁
DRP-SLDA	69.78	68.10	65.88	86.49	84.90	83.77
DRP-SLDA_NR	69.39	67.71	65.48	86.27	84.61	83.56
DRP-SLDA_NS	68.76	67.53	64.94	86.14	84.45	83.38
DRP-SLDA_ND	68.97	67.15	64.97	83.43	80.96	79.94
DRP-SLDA_NDR	67.31	65.94	63.27	83.07	80.34	79.37
DRP-SLDA_NDS	66.21	65.20	62.31	82.63	80.08	79.09

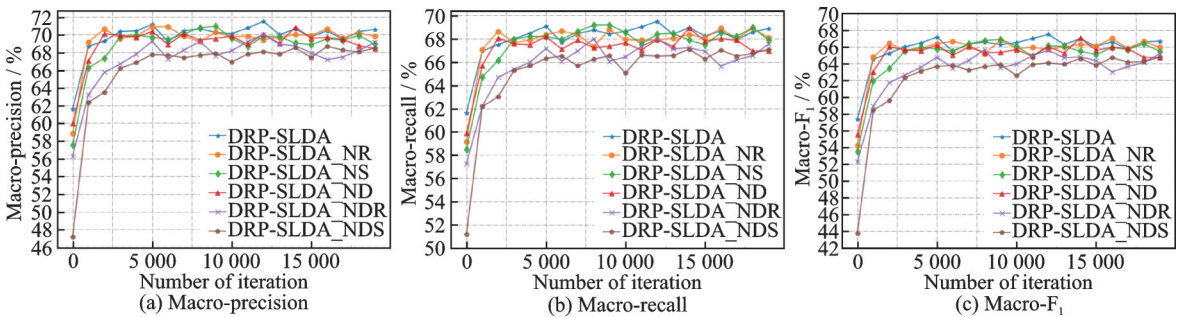


图 6 Sogou 数据集在 3-way 1-shot 下消融方法分类结果

Fig.6 Sogou data set classification results of ablation methods under 3-way 1-shot

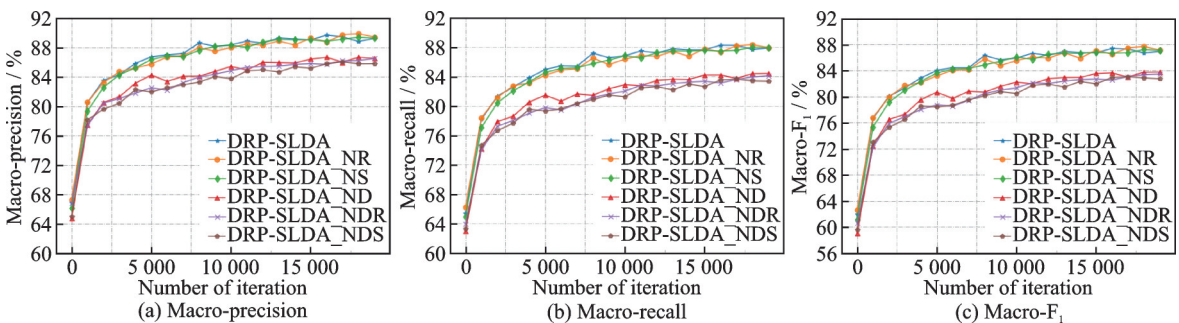


图 7 20newsgroup 数据集在 5-way 5-shot 下消融方法分类结果

Fig.7 20newsgroup data set classification results of ablation methods under 5-way 5-shot

如图6所示,在Sogou数据集上对于3-way 1-shot分类任务各消融方法结果,可以看出:(1)DRP-SLDA对比DRP-SLDA_NR方法,各分类指标提高0.39%、0.39%和0.40%,表明源集通用性特征表示的有效性,引入源集对模型有积极影响。(2)DRP-SLDA对比DRP-SLDA_NS方法,各分类指标提升1.02%、0.57%和0.94%,表明利用SLDA模型能有效增强词分布特征。(3)DRP-SLDA对比DRP-SLDA_ND方法,各分类指标提升0.81%、0.95%和0.91%,表明利用支持集和查询集样本,能提取语义增强的交叉特征 ψ 。

如图7所示,在20 newsgroup数据集上对于5-way 5-shot任务各消融方法结果,可以看出:对比DRP-SLDA_NDR方法,DRP-SLDA方法各分类指标降低3.06%、4.56%和4.40%;对比DRP-SLDA_NDS方法,DRP-SLDA方法各分类指标降低3.86%、4.82%和4.68%。表明在没有获取样本交叉特征时,SLDA模型提取的特定类词汇特征有助于提升DRP-SLDA模型泛化性能,而源集通过扩充数据样本提取词汇通用性特征对模型效果微效。

图8是对20newsgroup数据集的样本分布PCA降维,将样本映射到二维空间,可视化各消融方法的影响。图8(a)中3个聚类分簇显著,分类边界明显优于其他方法,说明DRP-SLDA模型生成具有类别区分性的样本表示。

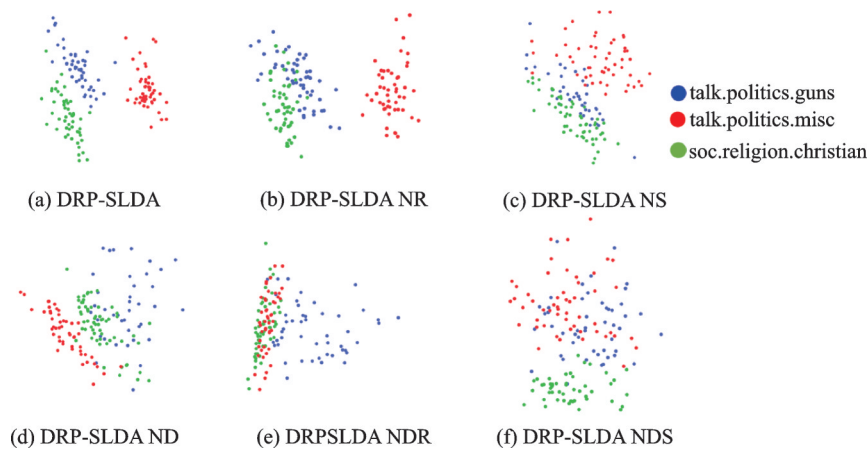


图8 消融方法在20newsgroup数据集的PCA可视化比较

Fig. 8 PCA visualization comparison of ablation methods in 20newsgroup data set

2.2.3 动态路由算法的有效性分析

为了验证DR-Proto网络中动态路由算法对DRP-SLDA模型分类效果的影响,在FewRel数据集上进行5-way实验,动态路由算法的随不同迭代次数的可视化如图9所示,展示在FewRel数据集上提取样本交叉特征 ψ 可视化结果。由图可知,随着迭代次数的增加,模型分类边界清晰。表明DRP-SLDA模型通过动态路由算法的多次迭代,能够有效提取样本交叉特征,动态获得使类别边界更清晰的原型。综上所述,所提出的DRP-SLDA模型能有效增强小样本文本分类的特征表示,提升原型的类别辨识度。

3 结束语

本文提出一种基于SLDA和动态路由的原型网络模型DRP-SLDA,利用SLDA模型获得词汇-类别的语义映射增强词的分布特征,结合动态路由算法更新样本权重为不同样本赋予权重获得动态原型,从而有效提升模型的泛化性能。在多个数据集上的对比实验表明了DRP-SLDA模型的有效性。下一步将对多标签小样本元学习方法展开研究。

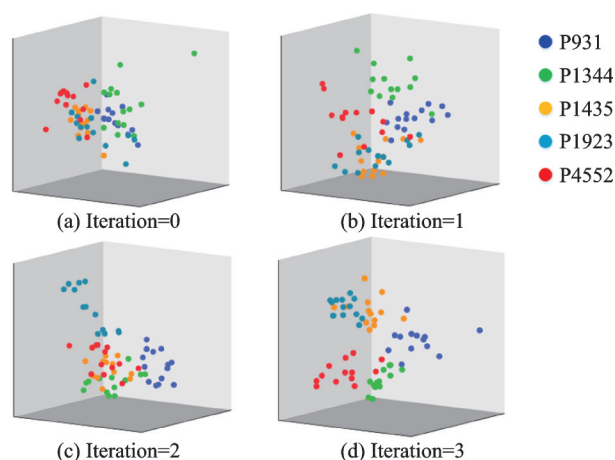


图9 不同动态路由迭代次数的样本(特征)可视化结果

Fig.9 Visualization results of samples (features) of different dynamic routing iteration times

参考文献:

- [1] KUANG Shaohui, LI Junhui, BRANCO A, et al. Attention focusing for neural machine translation by bridging source and target embedding [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics, 2018: 1767-1776.
- [2] ZHOU Chunting, SUN Chonglin, LIU Zhiyuan, et al. A C-LSTM neural network for text classification[J]. Computer Science, 2015, 1(4): 39-44.
- [3] YANG Zichao, YANG Diyi, DYER C, et al. Hierarchical attention networks for document classification [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016: 1480-1489.
- [4] NORBERT J, WLODZISLAW D, KRZYSZTOF G. Meta-learning in Computational Intelligence [M]. Berlin, Germany: Springer, 2011: 97-115.
- [5] LAKE B L, SALAKHUTDINOV R, TENENBAUM J B. One-shot learning by inverting a compositional causal process [C]//Proceedings of Advances in Neural Information Processing Systems. United States: Lake Tahoe, 2013: 2526-2534.
- [6] BENGIO Y, GUYON G, DROR V, et al. Deep learning of representations for unsupervised and transfer learning [J]. Workshop on Unsupervised & Transfer Learning, 2012: 17-36.
- [7] HOCHREITER S, YOUNGER A S, CONWELL P R. Learning to learn using gradient descent [C]//Proceedings of International Conference on Artificial Neural Networks. Berlin, Heidelberg: Springer, 2001: 87-94.
- [8] HUISAN M, VAN R J N, PLAAT A. A survey of deep meta-learning[J]. Artificial Intelligence Review, 2021, 54(6): 4483-4541.
- [9] VINYALS O, BLUNDELL C, LILICRAP T, et al. Matching networks for one shot learning [C]//Proceedings of Advances in Neural Information Processing Systems. Montreal, Quebec Canada: Morgan Kaufmann, 2016: 3637-3645.
- [10] VILALTA R, DRISSI Y. A perspective view and survey of meta-learning [J]. Artificial Intelligence Review, 2002, 18(2): 77-95.
- [11] SNELL J, KEVIN S, RICHARD Z. Prototypical networks for few-shot learning [C]//Proceedings of Advances in Neural Information Processing Systems. Montreal, Quebec Canada: Morgan Kaufmann, 2017: 4077-4087.
- [12] FORT S. Gaussian prototypical networks for few shot learning on omniglot [C]//Proceedings of International Conference on Learning Representations. Massachusetts, USA: OpenReview.net, 2018. DOI:10.48550/arXiv.1708.02735.

- [13] GAO Tianyu, HAN Xu, LIU Zhiyuan, et al. Hybrid attention-based prototypical networks for noisy few-shot relation classification [C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l]: AAAI, 2019: 6407-6414.
- [14] SUN Shengli, SUN Qingfeng, ZHON K, et al. Hierarchical attention prototypical networks for few-shot text classification [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, 2019: 476-485.
- [15] 唐焕玲, 窦全胜, 于立萍, 等. 有监督主题模型SLDA-TC文本分类新方法 [J]. 电子学报, 2019, 47(6): 1300-1308.
TANG Huanling, DOU Quansheng, YU Liping, et al. SLDA-TC: A novel text categorization approach based on supervised topic model[J]. Acta Electronica Sinica, 2019, 47(6): 1300-1308.
- [16] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules [C]//Proceedings of Advances in Neural Information Processing Systems. Long Beach, CA, USA: Curran Associates Inc, 2017: 3856-3866.
- [17] BAO Yujia, WU Menghua, CHANG Shiyu, et al. Few-shot text classification with distributional signatures [D]. Boston, Cambridge, USA: Massachusetts Institute of Technology, 2020.
- [18] SEPP H, JURGEN S. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
- [19] ARORA S, LIANG Yingyu, MA Tengyu. A simple but tough-to-beat baseline for sentence embeddings [C]//Proceedings of International Conference on Learning Representations. Massachusetts, USA: Open Review. net, 2017.
- [20] GENG Ruiying, LI Binhua, LI Yongbin, et al. Induction networks for few-shot text classification [C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, 2019: 3902-3911.
- [21] HINTON G E, KRIZHEVSKY A, WANG S D. Transforming auto-encoders [C]//Proceedings of Artificial Neural Networks & Machine Learning-International Conference on Artificial Neural Networks. Berlin, Heidelberg: Springer, 2011: 44-51.
- [22] CHEN Weiyu, LIU Yencheng, KIRA Z, et al. A closer look at few-shot classification [C]//Proceedings of In International Conference on Learning Representations. Massachusetts, USA: OpenReview. net, 2019.
- [23] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks [C]//Proceedings of the 34th International Conference on Machine Learning. Australia: PMLR, 2017: 1126-1135.
- [24] BERTINETTO L, HENRIQUES J F, TORR P, et al. Meta-learning with differentiable closed-form solvers [C]//Proceedings of In International Conference on Learning Representations. Massachusetts, USA: OpenReview. net, 2019.

作者简介:



张淑芳(1996-),女,硕士研究生,研究方向:数据挖掘, E-mail:1347604075@qq.com。



唐焕玲(1970-),通信作者,女,硕士生导师,研究方向:机器学习、人工智能、数据挖掘, E-mail:thl01@163.com。



郑涵(1997-),男,硕士研究生,研究方向:机器学习、数据挖掘, E-mail:zhenghan0503@163.com。



刘孝炎(1997-),男,硕士研究生,研究方向:机器学习, E-mail:lxyl5058247683@aliyun.com。



窦全胜(1971-),男,硕士生导师,研究方向:人工智能、机器学习、演化计算, E-mail:li_dou@163.com。



鲁明羽(1963-),男,教授,博士生导师,研究方向:机器学习、人工智能、数据挖掘, E-mail:lumingyu@dlmu.edu.cn。

(编辑:刘彦东)