文章编号：1674-2974（2023）02-0086-09 DOI：10.16339/j.cnki.hdxbzkb.2023269

基于域分离网络的实体解析迁移方法

孙琛琛[1†]，许雷[1]，申德荣[2]，聂铁铮[2]
（1. 天津理工大学 计算机科学与工程学院，天津 300384；
2. 东北大学 计算机科学与工程学院，辽宁 沈阳 110169）

摘 要：实体解析致力于识别多条记录是否描述真实世界相同实体，这是数据清洗和数据集成中的关键问题.近年来，基于深度学习的实体解析广受欢迎，它们需要大量标注数据才能达到较优的效果.然而，在现实场景中，大量高质量标注数据不容易获得.本文提出了一个基于深度迁移学习的实体解析模型，通过域分离网络提取源域和目标域的公共特征，并利用公共特征得到实体解析结果，从而实现从源域到目标域的迁移.实验结果表明，在多个数据集上，本文提出的方法比之前最好的方法在F_1度量上最大提高了40%左右.实验证明本文的方法具有更好的表现，并且训练时间更短.

关键词：实体解析；域分离网络；变分自编码器；数据集成；迁移学习
中图分类号：TP31 **文献标志码**：A

Domain Separation Network Based Entity Resolution Transferring Method

SUN Chenchen[1†], XU Lei[1], SHEN Derong[2], NIE Tiezheng[2]
(1. School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China;
2. School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)

Abstract: Entity resolution (ER) is a task to identify whether several records correspond to the same entity in the real world, which is a key problem in data cleaning and data integration. Recently, deep learning-based entity resolution is popular, which requires a large number of labeled data to achieve better results. However, a large number of high-quality labeled data are not always easily available in the real scenario. This paper proposes a deep transfer learning-based entity resolution model. The common features of the source domain and the target domain are extracted through a domain separation network. ER results are obtained by utilizing these common features. Therefore, the common features are transferred from the source domain to the target domain. The experimental results show that, on several datasets, the proposed method has a maximum improvement of about 40% in the F_1 metric compared with the previous best method. Experiments show that the proposed method has superior performance and shorter training time.

Key words: entity resolution; domain separation networks; variational auto-encoder; data integration; transfer learning

---

* **收稿日期**：2022-05-29
基金项目：国家自然科学基金资助项目（62002262，62172082，62072086，62072084），National Natural Science Foundation of China (62002262, 62172082, 62072086, 62072084)
作者简介：孙琛琛（1987—），男，山西晋中人，天津理工大学讲师，博士
† 通信联系人，E-mail：suncc@email.tjut.edu.cn

实体解析(Entity Resolution, ER)对重复数据删除、记录链接等应用和研究有着巨大的影响,它是数据清洗和数据集成中的一个基本问题[1].它在数据库、自然语言理解等多个领域都有巨大的应用价值.目前实体解析受到了很大的关注,但是还没有令人满意的解决方案.

实体解析[1]是给定可能具有错误、遗漏的记录对集合,它识别出引用现实世界同一实体的记录对.在过去几年里,深度学习技术获得了极大的进步.深度学习技术在文本、图像和语音等领域都获得了巨大的成功,深度学习使用带标签的数据,学习重要特征,减轻了昂贵的手工创建规则的负担,大大提高了学习的效率.深度学习目前成为替代传统机器学习的一种可行算法,例如支持向量机和决策树等,使用深度学习方法解决实体解析问题已成为当前的研究热点.

目前,只有当有大量标记的训练数据可用时,深度学习在性能上才会有较大的优势.但遗憾的是,现实中有时没有大量的带标签数据来进行学习.并且,许多深度学习方法只有在一个共同的假设下才能很好地工作——训练数据和测试数据来自相同的特征空间和分布.当分布发生变化时,大多数方法需要使用新收集的训练数据从头开始学习.在许多现实应用中,收集所需的训练数据并重建模型是费时费力的,甚至是不可能的.在这种情况下,使用迁移学习[2]将是必要的.

迁移学习是解决上述问题的一种很有效的方法,迁移学习关注的是知识或者特征的跨域迁移[2].一个人通过概括他的经验,就有可能实现从一种知识到另一种知识的转移.在生活中,学习国际象棋时,学过中国象棋的人可以比其他人学得更快,因为中国象棋和国际象棋之间可能有一些共同之处.受人类跨域传递知识能力的启发,迁移学习的目的是利用来自相关领域(称为源域)的知识来提高学习性能或最小化目标域中所需的标记样例的数量[2].迁移学习是利用数据、参数或领域之间的相似性,将在旧领域学习过的模型,应用于新领域的一种学习过程.迁移学习的关键是找到源域和目标域之间的相似性,并加以利用.

以往的实体解析方法大多假设存在大量的带标签数据供我们使用,因此,模型可以训练出较好的结果.但当不存在带标签数据供我们使用时,依靠大量标签数据才能得到较好结果的方法则无法使用,所以在实体解析领域应用迁移学习是很有必要的.本文假设本领域没有标签数据,而相似领域存在标签数据,在这种情况下提出方法,从而解决本领域因无法得到标签数据导致无法训练的问题.以往的方法,如 Kasai 等人[3]提出的低资源实体解析方法的网络结构较为简单,无法得到较好的迁移效果;Bogatu 等人[4]提出的变分主动实体解析模型(Variational Active Entity Resolution, VAER)方法需要先在其他领域进行预训练并且在本领域进行微调才能使用.在实体解析领域进行迁移学习首先需要得到记录中属性的向量,才能汇总得到记录的向量结果.在实体解析模型中,模型得到的是一对记录之间的关系,而在实体解析领域迁移,要迁移的是每条记录的知识或参数,而不是迁移一对记录之间的相似性,因此要精心设计迁移模型.

本文提出了使用域分类网络的基于深度学习的迁移方法,用于学习域不变表示.设计了用于实体解析的实体解析模型,具体地说,首先使用编码层中的编码器对属性信息进行编码得到属性的分布向量,然后将各自的属性分布向量,送入比较层,计算对应属性间的差异,得到记录间的比较结果,最后将比较结果向量送入分类器,得到最终的分类结果.随后将实体解析模型作为组件设计了基于域分离网络的迁移模型,迁移模型所用的编码器结构与实体解析模型编码层中编码器结构一致.利用域分离网络的编码器将域的私有特征和共享特征分离开来,再利用分离出的域共享特征进行分类,得到分类结果,从而达到从源域迁移到目标域的目的.其中分离出来的域共享特征为源域和目标域共享的特征.具体来说,本文的贡献如下:

1) 构造了一个深度实体解析模型,对属性进行编码,随后计算属性的相似性,最后对记录分类.实体解析模型训练速度较快,因此作为随后迁移模型的组件,迁移模型整体训练时间较短.

2) 将域分离网络应用到实体解析领域,提出了一种基于深度实体解析的迁移方法,该方法学习域私有表示和域共享表示,利用域共享表示达到域迁移的效果.

3) 在多个数据集上进行了实验评估,测试本文提出方法的有效性,进行了消融实验等,证明了提出的迁移方法的有效性.

本文第一节将介绍实体解析和迁移学习的相关工作.第二节介绍用于实体解析任务的匹配模型.第

三节提出了基于域分离网络的迁移学习方法.第四节介绍了所做的实验以及实验设置和数据集等细节,并在第五节进行了总结.

1 相关工作

1.1 实体解析

实体解析早期的研究工作致力于设计各种基于字符串的距离函数来度量成对记录的相似性.显然,这种无人监督的方法缺乏有效性和普遍性,并不存在针对所有数据集的单一度量.为了克服上述限制,基于机器学习的技术变得流行起来.这些方法将实体解析问题视为一个二分类任务,并将传统的分类器(如贝叶斯分类器)应用于手工制作的特征.它们可以在一定程度上提高实体解析的精度,但对人工特征工程的依赖仍然阻碍了通用性和鲁棒性.

目前,使用深度学习解决实体解析问题越来越受到欢迎,通过设计有效的深度学习模型来进一步提高性能.在给定一对文本记录的情况下,DeepER采用GloVe进行单词嵌入,然后应用长短期记忆网络(Long Short-Term Memory, LSTM)模型对实例的文本描述进行编码,随后进行分类训练[5].DeepMatcher使用注意力机制扩展了循环神经网络(Recurrent Neural Networks, RNNs),用于文本实例之间的实体匹配,将从每个属性导出的相似性向量连接起来,以形成分类器的输入[6].对于异构的记录或者记录内存在缺失、错误拼写、遗漏等情形,提出HierMatcher方法,从单词(token)层面对记录进行比较,有效避免脏数据集的影响[7].MCA (Multi-Context Attention)提出使用多种注意力,计算记录内和记录间的注意力,利用多种注意力关系进行匹配[8].

1.2 迁移学习

迁移学习首先由香港科技大学教授杨强提出.迁移学习允许训练和测试中使用的任务或者分布有所不同.迁移学习近年来受到越来越多的关注,大量的迁移学习相关方法被提出.跨域误差最小化(Cross-Domain Error Minimization, CDEM)方法旨在学习域不变特征,为此方法使用跨域误差最小化、源域和目标域分类误差最小化、分布对齐和鉴别性学习四个目标来保证学习域不变特征[9].统一联合分布对齐域自适应(Domain Adaptation with Unified Joint Distribution Alignment, UJDA)方法进行域和类两个级别的对齐,使用两个联合分类器并利用联合对抗性损失进行域自适应[10].跨域梯度差异最小化(Cross-Domain Gradient Discrepancy Minimization, CGDM)明确地将源样本和目标样本产生的梯度差异最小化,以实现类级别更好的分布对齐[11].特定域对抗网络(Domain-Specific Adversarial Network, DSAN)提出了同时输入域特征和域特殊信息到单一的编码器(Encoder)来学习不变表示的方法[12].语义集中域适应(Semantic Concentration for Domain Adaptation, SCDA)方法在分类器和特征提取器间对样本的预测分布进行对抗学习,从而获得对齐良好的特征[13].但是以上迁移学习方法并不适用于实体解析这样需要学习记录对间相互关系的领域.域对抗神经网络(Domain-Adversarial Neural Network, DANN)通过在域分类器和特征提取器间加入梯度反转层训练模型,达到混淆域分类器的目的[14].域分离网络(Domain Separation Networks, DSN)方法利用编码器和解码器提取域的私有特征和共享特征,分类器对共享编码器的输出分类,得到分类结果[15].

2 深度实体解析模型

本节介绍用于实体解析的模型,此模型将作为组件用于下一节将要介绍的实体解析迁移模型.模型学习各个属性间的相似性,并进行实体匹配.具体地说,给定均由属性$A_1,\cdots,A_m$组成的一对记录(l_1,l_2),属性可以视为由单词组成的序列.实体解析的目标是判断l_1和l_2是否属于真实世界的同一实体.表1列出了三条实体记录示例,分别由三个属性组成一条记录.其中R_1和R_2是匹配的,R_1和R_3是不匹配的.

图1给出了实体解析的模型.给定一对记录(l_1,l_2),首先使用词嵌入fastText[16]为记录中的每个属性生成嵌入序列($\boldsymbol{w}_1,\cdots,\boldsymbol{w}_m$),其中$\boldsymbol{w}_1$为属性$A_1$对应的属性嵌入向量序列.接下来每个属性嵌入序列通过双向门控循环单元(Gated Recurrent Unit, GRU)捕获序列内的上下文关系生成各自的属性嵌入向量($\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m$).在记录对的每个属性经过编码器生成属性嵌入后,记录对的对应属性间进行相似性比较,将相似性比较的结果作为分类器的输入,分类器输出最终的匹配概率.

2.1 输入层

输入层主要用于得到记录中属性的上下文信息,分别得到各个属性向量表示.因此,在给定文本的记录对时,首先要将文本转换为嵌入向量,相比于Word2Vec和GloVe,fastText在处理词典外单词方面具有一定的优势,因此本文使用fastText.文本转换为

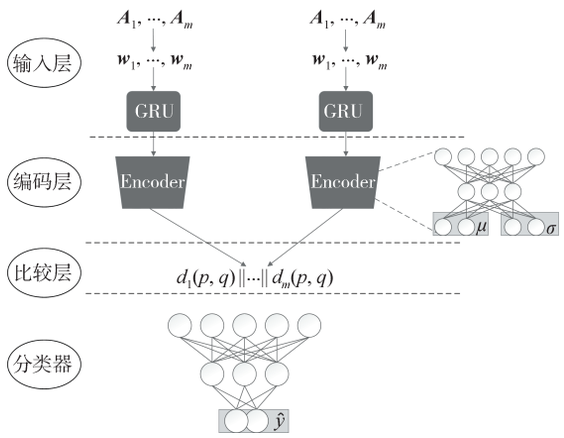

图 1　深度实体解析框架

Fig.1　Framework of deep entity resolution

表 1　实体解析示例

Tab.1　Example of entity resolution

ID	Title	Manufacturer	Price
R_1	adobe photoshop cs3 [mac]	adobe	649
R_2	adobe photoshop cs3 for mac	adobe	609.99
R_3	adobe cs3 after effects	adobe	1 025.99

嵌入后,要想得到各自的属性向量,一种方法是属性里的单词嵌入序列相加求均值作为各自的属性向量表示,但是这种方法不能很好地提取单词之间的上下文关系;另一种方法是利用循环神经网络模型得到属性向量,该方法能捕获单词间的上下文关系,因此选用此方法.

给定文本记录对(l_1,l_2)作为输入,首先将输入的记录矢量化,使用fastText词嵌入得到输入记录的矢量化表示$(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_m)$,其中每个记录的每个属性均为词嵌入序列.得到记录的矢量化表示后,接下来生成各个属性的嵌入表示.使用一个双向的GRU接收每个属性序列,利用GRU对属性序列编码得到属性向量$(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m)$,并将记录对的属性向量集合分别送入属性编码器.通过对比发现,采用门控结构的GRU模型比长短期神经网络模型和递归神经网络模型表现更好.如公式(1)所示.

$$\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m=\mathrm{BiGRU}(\boldsymbol{w}_1,\cdots,\boldsymbol{w}_m) \tag{1}$$

2.2　属性编码层

属性编码层用于对得到的属性向量进一步的编码和压缩并得到属性隐含的分布式表示.本文利用变分自编码器(Variational Auto-Encoders, VAE)[17]的结构,利用VAE的编码器对属性嵌入序列编码,VAE在提取深层的掩藏表示和重建方面具有一定的优势,利用VAE对属性向量提取分布式表示.VAE的编码器用于生成均值μ和方差σ.VAE的编码器生成的μ和方差σ一起描述了属性的分布,即(μ,σ)代表了一个属性,因此属性编码层生成的分布可以用于下一层属性比较层的计算.

属性编码层由两个参数共享的属性编码器组成,输入层用循环神经网络模型得到了属性的向量表示,两个编码器分别将属性向量集合作为输入,并通过带有非线性激活函数的一到多个线性层.对于每个属性向量表示,利用对角协方差(μ,σ)拟合潜在的高斯分布.每个编码器分别生成实体表示$\{(\mu_1,\sigma_1),\cdots,(\mu_m,\sigma_m)\}$,每个属性值对应一个$(\mu,\sigma)$.两条记录的比较通过对应属性值生成的分布来计算.属性编码层利用对角协方差拟合属性嵌入向量的分布,利用分布间的距离判断对应属性间是否相似,公式如下:

$$(\mu_1,\sigma_1),\cdots,(\mu_m,\sigma_m)=\mathrm{Encoder}(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m) \tag{2}$$

2.3　属性比较层

属性比较层计算对应属性间的相似性,由于属性编码器输出为高斯分布,对于量化两个高斯分布间的距离可以使用Wasserstein距离[18].d-Wasserstein距离描述了当成本由L^d距离给定时,将一个概率测度的单位质量传输到另一个概率测度的单位质量的最小成本[18].在本文中,使用d=2的平方后的Wasserstein距离(W^2)计算属性相似性.例如,如果计算两个k维对角高斯分布p和q之间的W^2距离,公式如下:

$$W^2(p,q)=\sum_i^k=1(\mu_i^p-\mu_i^q)+(\sigma_i^p-\sigma_i^q) \tag{3}$$

属性比较层用于比较属性间的相似度,并将比较后的结果送入实体解析分类器.Wasserstein距离用于计算两个概率之间的距离,因此可以用于计算属性编码层输出分布之间的距离.当两个属性编码器输出$\{(\mu_1,\sigma_1),\cdots,(\mu_m,\sigma_m)\}$和$\{(\mu_1',\sigma_1'),\cdots,(\mu_m',\sigma_m')\}$到属性比较层,计算$m$个对应属性间的Wasserstein距离向量$\boldsymbol{d}_{\mathrm{w}}=(\mu-\mu')^2+(\sigma-\sigma')^2$.最后,将$m$个计算出的向量拼接起来,送入实体解析分类器.

2.4　实体解析分类器

实体解析分类器区分一对记录是否为同一实体.实体解析分类器接收上一层传入的m个拼接起来的距离向量并送入到两层具有非线性激活函数的多层感知器(Multilayer Perceptron, MLP)中,再将线性层的输出经过Softmax函数得到归一化输出,将其分类为匹配或不匹配.公式如(4)所示,其中$\boldsymbol{d}_{\mathrm{w}}$代表$m$个对应属性分布计算后的距离向量拼接后的向量,ReLU为激活函数.

$$\hat{y}=\mathrm{Softmax}(\mathrm{ReLU}(\mathrm{MLP}(\boldsymbol{d}_{\mathrm{w}}))) \tag{4}$$

实体解析任务优化目标是最小化分类器的分类误差.其中L_c代表损失函数,y为真实的标签,$\hat{y}$表示经过实体解析分类器后输出的预测标签.定义损失函数如下:

$$L_c=y\log(\hat{y})+(1-y)\log(1-\hat{y}) \tag{5}$$

3 深度实体解析迁移模型

在迁移环境下,本文改进域分离网络结构适应实体解析任务进行迁移学习.在给定源域数据集带标签、目标域数据集不带标签的情况下,本文的训练目标是利用源域和目标域的数据使模型能准确预测目标域数据的标签.定义源域$\boldsymbol{X}_S$,其中有N_S个带标签的数据,目标域$\boldsymbol{X}_T$,其中有N_t个不带标签的数据.本文假设源域和目标域的记录共享相同的属性模式.域分离网络显式建模了域的私有表示和域的共享表示.域分离网络利用不同损失函数的组合实现了源域和目标域分别有一个域私有表示,同时源域和目标域有一个域共享表示.分类器通过对域共享表示部分的输出进行分类得到分类结果.对连接域共享表示的分类器进行分类,能更好地跨域泛化,不受域私有表示的影响.本文利用变分主动实体解析模型(VAER)作为提取特征的基本组件,结合域分离网络(DSN)的思想,提出了实体解析迁移模型VAERDSN.

将源域和目标域记录对集合中所有记录嵌入得到每条记录的属性嵌入序列,源域记录对嵌入集合为$\boldsymbol{X}_S$,目标域记录对嵌入集合为$\boldsymbol{X}_T$.如图2所示,$\boldsymbol{X}_S$和$\boldsymbol{X}_T$为VAERDSN的输入;$\boldsymbol{h}_o^s$代表$\boldsymbol{X}_S$经过GRU得到隐藏表示再输入到源域私有编码器E_o^s得到的源域私有表示向量;$\boldsymbol{h}_u^s$代表$\boldsymbol{X}_S$经过GRU得到隐藏表示再输入到共享编码器E_u得到的源域共享表示向量,$\boldsymbol{h}_u^t$和$\boldsymbol{h}_o^t$同理;$\hat{e}_1,\cdots,\hat{e}_m$表示将域私有表示和域共享表示经过解码器$D$的重建输出;$\hat{y}$为将源域共享表示输入到分类器$C$得到的预测标签;$d'_k$为域共享特征输入到域分类器得到的域预测标签;$L_c$、$L_{\text{difference}}$、$L_{\text{similarity}}$和$L_{\text{recon}}$为不同的损失函数.

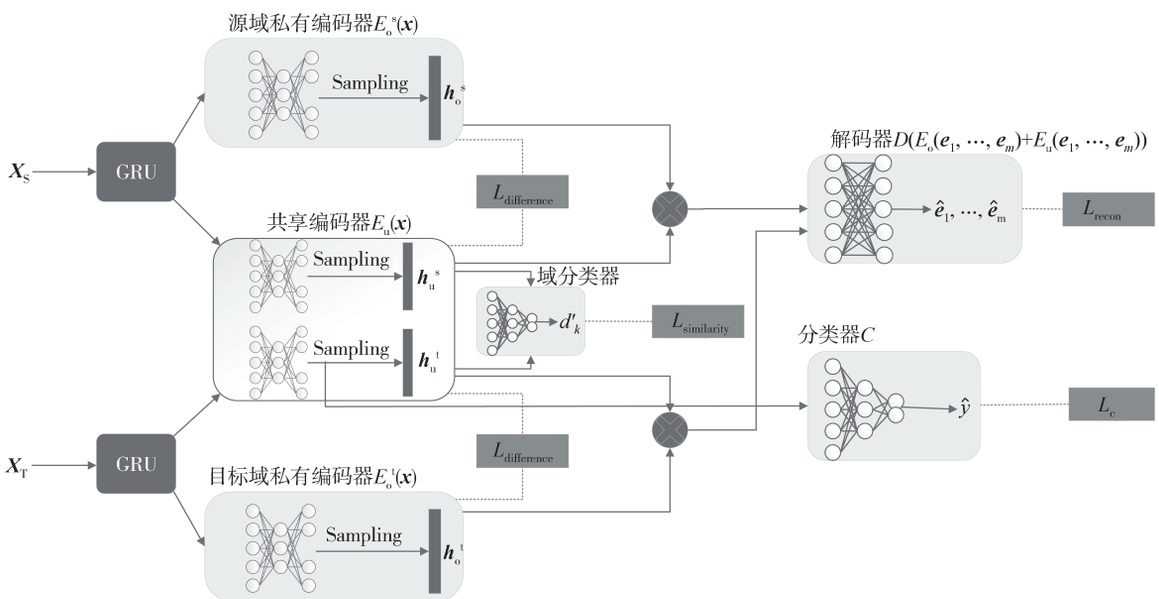

图2 迁移网络架构

Fig.2 Framework of the transfer learning network

3.1 迁移网络结构之编码器

迁移模型的编码器$E(\boldsymbol{x})$旨在提取属性的隐藏表示,为后续网络结构提供更好的特征.编码器$E(\boldsymbol{x})$分为私有编码器$E_o(\boldsymbol{x})$和共享编码器$E_u(\boldsymbol{x})$,两种编码器结构与实体解析任务中的属性编码器结构相同.其中私有编码器$E_o(\boldsymbol{x})$分为源域私有编码器$E_o^s(\boldsymbol{x})$和目标域私有编码器$E_o^t(\boldsymbol{x})$,负责提取域独有的特征;共享编码器$E_u(\boldsymbol{x})$负责提取源域和目标域公共的特征.编码器$E(\boldsymbol{x})$将特征转换为嵌入向量,为下一步的解码器提供输入.

因为源域和目标域的私有特征不同,因此需要两个编码器分别提取源域和目标域的私有特征.提取源域和目标域的公共特征时,可以只使用一个编码器来达到提取公共特征的目的.例如当送入源域数据时,源域数据经过源域私有特征编码器得到源域私有特征向量,源域数据经过域共享编码器得到域共享特征向量.

数据进入编码器中首先会生成属性分布表示

(μ,σ),属性分布经过Sampling操作后,得到对应的属性向量表示.在高斯分布(0,1)中采样ε,用Sampling公式$\boldsymbol{h}=\mu+\varepsilon\times\sigma$表示从属性分布到属性向量的变换.公式(6)中的Encoder即为实体解析模型中编码层的Encoder.数据经过编码器过程如下:

$$\boldsymbol{h}_u=\text{Sampling}(\text{Encoder}(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m)) \tag{6}$$

3.2 迁移网络结构之解码器

解码器$D(\boldsymbol{h})$将编码器的输出重建回属性表示.迁移模型输入源域数据时,解码器接收源域私有编码器$E_{\text{o}}^{\text{s}}(\boldsymbol{x})$输出的源域私有特征和共享编码器$E_{\text{u}}(\boldsymbol{x})$输出的源域和目标域共享特征相加得到的向量作为输入,经过解码器得到源域的重建属性表示.目标域数据工作方式与源域数据一致.解码器$D(\boldsymbol{h})$由两层带有非线性激活函数的多层感知器构成,输出用于重建损失.解码器$D(\boldsymbol{h})$的存在保证了编码器$E(\boldsymbol{x})$编码有效的特征,避免了编码器学习到与任务无关的参数,公式如下:

$$\hat{e}_1,\cdots,\hat{e}_m=D(E_{\text{o}}(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m)+E_{\text{u}}(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m)) \tag{7}$$

解码器用于保证解码器学习到的知识与任务相关.如果编码器的输出不经过解码器进行重构,编码器输出在极端情况下会跟编码器的输入相似,而本文想要让编码器学习输入的属性向量的隐藏分布.因此为了避免编码器不学习属性向量的隐藏表示,在编码器后面接入解码器,计算解码器的输出与编码器的输入之间的重构误差,保证编码器学习到的知识与任务有关.

3.3 迁移网络结构之分类器

分类器$C(\boldsymbol{h})$采用与实体解析任务中实体解析分类器相同的结构,由带有非线性激活函数的多层感知器构成,输入和输出与公式(4)相同.分类器对源域数据经过共享编码器$E_{\text{u}}(\boldsymbol{x})$的输出进行分类任务,得到最终分类结果.只使用共享编码器$E_{\text{u}}(\boldsymbol{x})$的输出进行分类可以减少域私有特征的影响,只使用公共特征进行分类,从而得到更好的迁移到目标域的效果.分类器会输出最终的分类结果,由于在训练时,目标域数据没有标签,因此只有源域数据会经过分类器输出分类结果.

3.4 损失函数

本小节介绍迁移模型的损失函数.迁移模型训练目标是将总损失L降到最低,公式如下:

$$L=L_{\text{c}}+\alpha L_{\text{difference}}+\beta L_{\text{similarity}}+\gamma L_{\text{recon}} \tag{8}$$

其中α、β和γ均为控制损失项的超参数;L_{c}为分类任务损失;$L_{\text{difference}}$为差异性损失,保证域私有特征和域共享特征之间的差异性;$L_{\text{similarity}}$为相似性损失,保证源域和目标域各自提取的共享特征相似;最后,L_{recon}表示重建损失,希望重建回的属性表示与编码器的输入一致.

分类任务损失L_{c}表示模型预测标签的能力,希望损失越小越好,它同实体解析模型损失函数定义一致.因为目标域没有标签,因此只有带标签的源域数据经过分类器.

差异性损失在源域的私有特征和共享特征或目标域的私有特征和共享特征之间计算,差异性损失保证了私有编码器$E_{\text{o}}(\boldsymbol{x})$和共享编码器$E_{\text{u}}(\boldsymbol{x})$之间提取记录的不同方面.差异性损失利用Wasserstein距离对私有编码器和共享编码器各自输出的属性分布表示.因为目标域计算差异性损失和源域一致,因此只介绍源域计算差异性损失.具体地说,给定$\boldsymbol{X}_{\text{S}}$作为输入后,源域私有编码器和共享编码器的输出均为记录对的分布表示,分别为$(\mu_i^{\text{o}},\sigma_i^{\text{o}})$和$(\mu_j^{\text{u}},\sigma_j^{\text{u}})$,其中$i$和$j$分别对应第$i$和$j$个属性,$i,j\in[1,m]$.接下来计算两个编码器输出的对应属性分布间的距离,即用Wasserstein距离计算当i和j相等时$(\mu_i^{\text{o}},\sigma_i^{\text{o}})$和$(\mu_j^{\text{u}},\sigma_j^{\text{u}})$之间的距离.差异性损失定义如下:

$$L_{\text{difference}}=\sum_{k=0}^{N_{\text{s}}}\sum_{i,j=1}^{m}(\mu_i^{\text{o}}-\mu_j^{\text{u}})^2+(\sigma_i^{\text{o}}-\sigma_j^{\text{u}})^2 \tag{9}$$

相似性损失鼓励源域数据和目标域数据经过共享编码器后的表示尽可能相似,而与域无关.使用域对抗相似性损失来训练模型,迷惑域分类器使之不能正确地判断数据来自源域或目标域.相似性损失通过梯度反转训练域共享编码器学习域无关的特征,实现混淆域分类器的作用.其中,d_k是样本k的真实域标签,d'_k是域分类器输出的样本k的预测域标签.相似性损失定义如下:

$$L_{\text{similarity}}=\sum_{k=0}^{N_{\text{s}}+Nt}d_k\log d'_k+(1-d_k)\log(1-d'_k) \tag{10}$$

重建损失对经过解码器的输出$(\hat{e}_1,\cdots,\hat{e}_m)$与编码器的输入$(\boldsymbol{e}_1,\cdots,\boldsymbol{e}_m)$进行比较,确保编码器的有效性.源域和目标域都要计算重建损失.其中$\boldsymbol{e}_b$为编码器的输入,$\hat{e}_b$为解码器的输出.对输入编码器的属性向量和解码器输出的重建属性向量按位相减并求均值.本文使用均方误差来计算重建损失,均方误差对两个向量按位相减:

$$L_{\text{recon}}=\frac{1}{m}\sum_{b=1}^{m}(e_b-\hat{e}_b)^2 \tag{11}$$

4 实验评估

本文使用两对共四个数据集进行了实验.其中所有数据集均已经过分块操作,每个数据集随机分为训练、验证、测试数据,比例为3∶1∶1.表2给出了数据集的统计数据,包含数据集大小、匹配大小、属性个数等.其中Zomato-Yelp(ZY)、Fodors-Zagats(FZ)[6]是餐馆数据集,Books3、Books4是书籍数据集.其中Zomato-Yelp、Books3和Books4数据集均来自AnHai's Group[6].超参数α设置为0.01,β设置为0.075,γ设置为0.25.系统全面地展示了本文提出方法的有效性.

表2 数据集

Tab.2 Data Set

数据集	领域	大小	匹配对数	属性
Fodors-Zagats	Restaurant	946	110	6
Zomato-Yelp	Restaurant	894	450	4
Books3	Book	450	134	4
Books4	Book	450	107	4

使用精确率(P)、召回率(R)和F_1分数作为实验的评价指标.精确率衡量预测集合中正确预测的比例,召回率衡量真实匹配集合中被正确预测的比例,F_1为两者的调和平均数.

为了评估实体匹配模型的有效性,使用DeepMatcher、MCA模型与本文实体解析模型比较.由于迁移模型使用了较多的编码器,因此时间性能较为重要,在分类效果类似的情况下,本文更关注时间性能.

对于每个目标数据集,源由另外一个数据集给出(例如,ZY的源是FZ).图3、图4展示了迁移模型的性能.在图3中,可以看到当FZ为源数据集,ZY为目标数据集时,在源数据集上训练出来的模型直接在目标数据集上测试时的F_1为66.92%,当使用了迁移模型后,F_1提高到了83.8%,提高了大约17%.使用本文模型在没有目标标签的情况下达到了较高水平.在ZY为源数据集、FZ为目标数据集时,使用迁移模型后,相比硬迁移F_1提高了大约5%.

在图4中,当Books3为源数据集,Books4为目标数据集时,在源数据集上训练出来的模型直接在目标数据集上测试时的F_1为37.35%,使用迁移模型后,F_1提高了不到1%.在Books4为源数据集、Books3为目标数据集时,使用了迁移模型后,相比于直接使用源数据集上训练出来的模型,F_1由71.2%提高到了86.92%.

在图3和图4中,可以观察到由一个数据集迁移到另一个数据集时存在难易程度不同的现象,例如,由Books3迁移到Books4,模型提升效果很小,但当由Books4迁移到Books3时,模型提升效果较大,可以认为是数据集间蕴含的语义信息有较大差异.

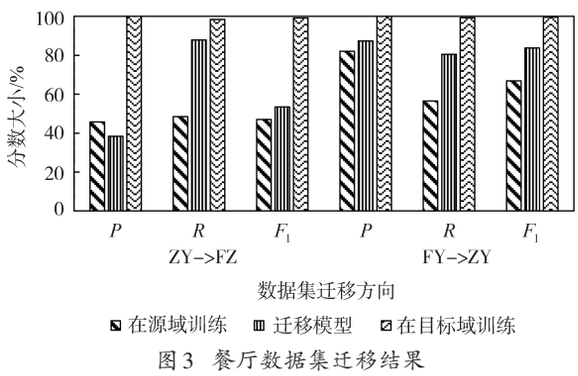

图3 餐厅数据集迁移结果

Fig.3 Transfer learning results on restaurants

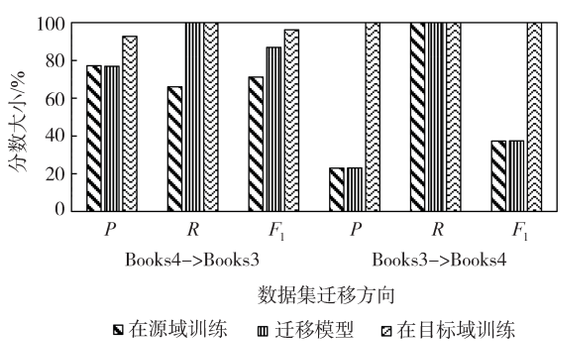

图4 书籍数据集迁移结果

Fig.4 Transfer learning results on books

本文与Kasai等人提出的适用于低资源实体解析方法进行了对比.如图5所示,除了Books4数据集,本文的方法VAERDSN均比对比的方法性能要好.在餐厅的两个数据集上,本文提出的方法相比Kasai等人提出的方法有了较大的提升,如在ZY数据集上,F_1由43.76%提升到了83.8%,有较大的提高.在Books3数据集上,本文的方法相比低资源模型方法,F_1由80.45%提高到86.92%.

接下来对比了实体解析模型与以前提出的深度实体解析方法在性能和时间上的差异.如图6所示,本文的实体解析模型在Books4数据集上的F_1超过了DeepMatcher但与MCA仍有差距.在另外三个数据集上,本文的实体解析模型与之前提出的两个方法有一定的差距,如在Books3数据集上,本文的实体解析

模型训练得出的F_1与另外两种方法分别差了大约8%和4%.如图7所示,本文的实体解析模型的训练效率高于DeepMatcher和MCA方法.在Books3数据集上,本文提出的方法和DeepMatcher训练时间相差达到了5倍之多.

因为迁移模型不使用解码器也能训练,因此本文进行了消融实验,其中VAERDSN-Decoder代表去掉解码器的模型.如图8所示,在去掉解码器之后,FZ和ZY数据集性能均有不同程度的下降.在FZ数据集上,F_1由53.46%降到了31.58%,在ZY数据集上F_1降低得最多,由83.8%降到了38.55%.由此可以看出迁移模型中,解码器保证了编码器提取的特征有利于迁移任务的进行,保证了编码器向有利于任务的方向训练.

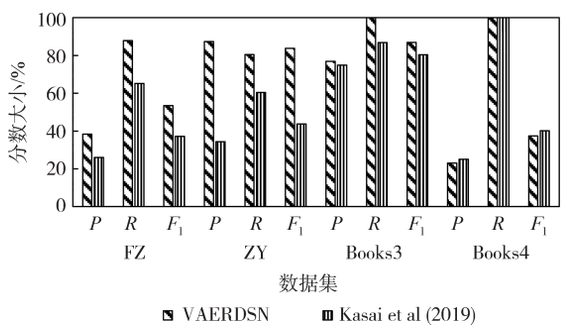

图5　迁移方法有效性

Fig.5　Effectiveness of transfer learning

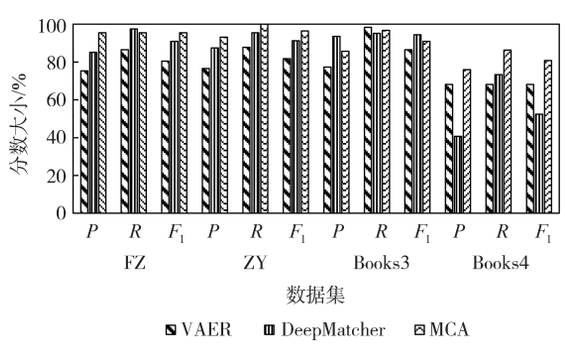

图6　实体解析模型性能对比

Fig.6　Performance comparison of entity resolution

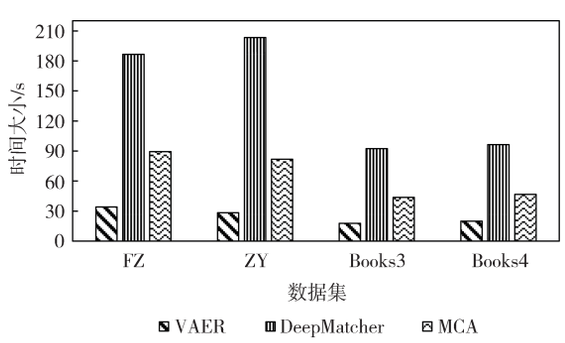

图7　实体解析训练时间对比

Fig.7　Comparison of entity resolution training time

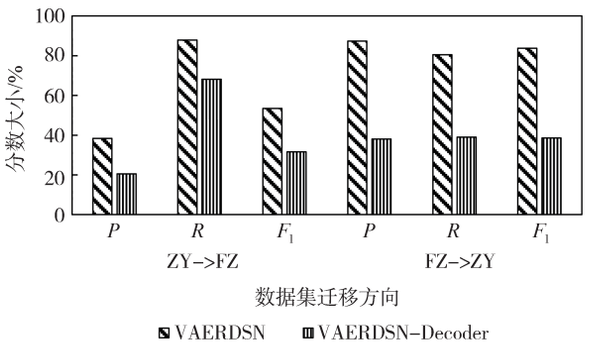

图8　解码器消融研究结果

Fig.8　Decoder ablation study results

本文还进行了参数实验,比较不同隐藏层大小对性能的影响.如图9所示,选取了ZY和Books3两个数据集进行展示.在ZY数据集上,可以看出选取隐藏层大小为100时,F_1最小,为80.43%;在隐藏层大小为200时,F_1达到了83.8%;隐藏层大小为250时,F_1为83.91%.隐藏层大小逐渐增大,F_1也在逐渐增高,且在隐藏层大小为200和250之间的F_1差距不大.在Books3数据集上,显示出了一样的规律,隐藏层大小由100增大到250,F_1也逐渐增大,并且在隐藏层大小选取200和250时,F_1差距不到0.5%.因此,本文设定隐藏层大小为200.

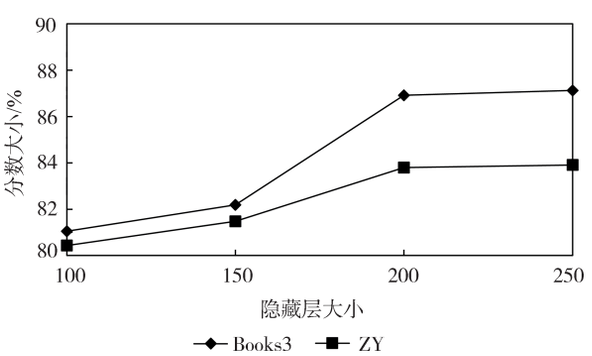

图9　隐藏层不同大小时F_1分数

Fig.9　F_1 score for different sizes of hidden layers

本文在多个数据集上进行了迁移实验以验证模型的有效性,迁移模型的实验结果相比于直接在源域训练均有不同程度提升.本文还将VAERDSN模型与其他迁移模型进行了对比,除了Books4数据集,其他数据集的实验结果都比所对比的模型结果要好.本文还将实体解析模型训练时间和其他实体解析模型训练时间进行了对比,可以看出,本文方法的训练时间要明显低于其他模型的训练时间.综上所述,本文通过多方面实验证明了VAERDSN的先进性.

5 结论

本文提出了基于深度学习和迁移学习的实体解析框架,解决实体解析领域的迁移学习.利用变分自编码器中的编码器结构作为组件,将编码器结合到域分离网络中,域分离网络利用不同功能的编码器提取出了域私有特征和域共享特征,并将域共享特征用作分类器的输入,其中解码器用于将提取出的特征重建回编码器的输入,编码器、解码器和分类器共同构成了本文提出的模型.通过实验证明了本文方法具有较好的迁移能力.通过展示迁移模型的性能,与以前提出的方法进行对比,以及利用参数研究,证明了本文提出方法的有效性.本文通过两对数据集的互相迁移,通过实验证明了本文提出的模型可以从源域和目标域中学习到公共知识并迁移到目标域中.

参考文献

[1] NAUMANN F, HERSCHEL M. An introduction to duplicate detection [J]. Synthesis Lectures on Data Management, 2010, 2 (1): 1-87.

[2] PAN S J, YANG Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[3] KASAI J, QIAN K, GURAJADA S, et al. Low-resource deep entity resolution with transfer and active learning [EB/OL]. 2019: arXiv:1906.08042[cs. DB]. https://arxiv. org/abs/1906.08042.

[4] BOGATU A, PATON N W, DOUTHWAITE M, et al. Cost - effective variational active entity resolution [C]//2021 IEEE 37th International Conference on Data Engineering. Chania, Greece: IEEE, 2021: 1272-1283.

[5] EBRAHEEM M, THIRUMURUGANATHAN S, JOTY S, et al. Distributed representations of tuples for entity resolution [J]. Proceedings of the VLDB Endowment, 2018, 11(11): 1454-1467.

[6] MUDGAL S, LI H, REKATSINAS T, et al. Deep learning for entity matching: a design space exploration [C]//SIGMOD '18: Proceedings of the 2018 International Conference on Management of Data. 2018: 19-34.

[7] FU C, HAN X P, HE J M, et al. Hierarchical matching network for heterogeneous entity resolution [C]//IJCAI. 2020: 3665-3671.

[8] ZHANG D X, NIE Y Y, WU S, et al. Multi-context attention for entity matching [C]//WWW '20: Proceedings of the Web Conference 2020.2020: 2634-2640.

[9] DU Y T, CHEN Y H, CUI F L, et al. Cross-domain error minimization for unsupervised domain adaptation [C]// DASFAA 2021: Database Systems for Advanced Applications.2021: 429-448

[10] DU Y T, TAN Z W, ZHANG X W, et al. Unsupervised domain adaptation with unified joint distribution alignment [C]// DASFAA 2021: Database Systems for Advanced Applications. 2021: 449-464.

[11] DU Z K, LI J J, SU H Z, et al. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 3936-3945.

[12] STOJANOV P, LI Z R, GONG M M, et al. Domain adaptation with invariant representation learning: what transformations to learn? [J]. Neural Information Processing Systems 2021.2021: 24791-24803.

[13] LI S, XIE M X, LV F R, et al. Semantic concentration for domain adaptation [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, 2021: 9082-9091.

[14] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks [EB/OL]. 2015: arXiv: 1505.07818 [stat. ML]. https://arxiv. org/abs/1505.07818.

[15] BOUSMALIS K, TRIGEORGIS G, SILBERMAN N, et al. Domain separation networks [EB/OL]. 2016: arXiv: 1608.06019 [cs. CV]. https://doi.org/10.48550/arXiv.1608.06019.

[16] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification [J]. 2016: arXiv: 1607.01759 [cs.CL]. https://doi.org/10.48550/arXiv.1607.01759.

[17] KINGMA D P, WELLING M. Auto-encoding variational bayes [J]. 2013: arXiv: 1312.6114 [stat.ML]. https://doi.org/10.48550/arXiv.1312.6114.

[18] MALLASTO A , FERAGEN A . Learning from uncertain curves: the 2-wasserstein metric for Gaussian processes [C]// Neural Information Processing Systems. NIPS Proceedings, 2017.