DOI:10.7544/issn1000-1239.20200693
59(3):674-682,2022

基于度量学习的无监督域适应方法及其在死亡风险预测上的应用

蔡德润　　李红燕
(北京大学信息科学技术学院　北京　100871)
(机器感知与智能教育部重点实验室(北京大学)　北京　100871)
(cdr@stu.pku.edu.cn)

A Metric Learning Based Unsupervised Domain Adaptation Method with Its Application on Mortality Prediction

Cai Derun and Li Hongyan
(*School of Electronics Engineering and Computer Science*, *Peking University*, *Beijing* 100871)
(*Key Laboratory of Machine Perception* (*Peking University*), *Ministry of Education*, *Beijing* 100871)

Abstract　Deep learning models have been widely used in the field of healthcare prediction tasks and have achieved good results in recent years. However, deep learning models often face the problems of insufficient labeled training data, the overall data distribution shift, and the category level data distribution shift, which leads to a decrease in the accuracy of the models. To solve the above problems, we propose an unsupervised domain adaptation method based on metric learning (additive margin softmax based adversarial domain adaptation, AMS-ADA). Firstly, this method uses the long short-term memory network with the attention mechanism to extract features. Secondly, this method introduces the idea of the generative adversarial network and reduces the overall data distribution shift via adversarial domain adaptation. Thirdly, this method introduces the idea of metric learning, which further reduces the category level data distribution shift by maximizing the decision boundary in the angular space. This method improves the effect of domain adaptation and the accuracy of the model. We perform the mortality prediction task of ICU patients in real-world healthcare datasets. The experimental results show that compared with other baseline models, our method can better solve the problem of data distribution shift and achieve better classification accuracy.

Key words　unsupervised domain adaptation; deep learning; mortality prediction; domain adversarial network; metric learning; attention mechanism

摘　要　近年来,深度学习模型已在医疗领域的预测任务上得到广泛应用,并取得了不错的效果.然而,深度学习模型常会面临带标签训练数据不足、整体数据分布偏移和类别之间数据分布偏移的问题,导致模型预测的准确度下降.为解决上述问题,提出一种基于域对抗和加性余弦间隔损失的无监督域适应方法(additive margin softmax based adversarial domain adaptation, AMS-ADA).首先,该方法使用带有注意力机制的双向长短程记忆网络来提取特征.其次,该方法引入了生成对抗网络的思想,以域对抗的形式减少了整体数据之间数据分布偏移.然后,该方法引入了度量学习的思想,以最大化角度空间内决策边界的方式进一步减少了类别之间的数据分布偏移.该方法能够提升域适应的效果与模型预测的

收稿日期:2020-09-07;**修回日期**:2021-02-08
基金项目:国家重点研发计划项目(2021YFE0205300);国家自然科学基金项目(62172018,62102008)
This work was supported by the National Key Research and Development Program of China (2021YFE0205300) and the National Natural Science Foundation of China (62172018, 62102008).
通信作者:李红燕(leehy@pku.edu.cn)

准确度.在真实世界的医疗数据集上进行了重症监护病人死亡风险预测任务,实验结果表明:由于该方法相较于其他5种基线模型能够更好地解决数据分布偏移的问题,取得比其他基线模型更好的分类效果.

关键词 无监督域适应;深度学习;死亡风险预测;域对抗网络;度量学习;注意力机制

中图法分类号 TP18; TP391

突发公共卫生安全事件往往会对社会医疗资源造成巨大的压力.例如2020年初新冠肺炎疫情的暴发所带来的医护人员人手短缺、医疗资源挤兑等问题.其原因之一是新型冠状病毒感染者容易出现"炎症风暴"[1],导致病情迅速恶化,死亡风险上升.医护人员需要投入大量的精力去观察和跟踪患者生理状况的变化,并需要根据患者死亡风险程度调配不同的医疗设备.例如体外膜肺氧合设备能够为抢救赢得宝贵的时间,但是数量比较少,适用于重症心肺功能衰竭患者.如果能够利用患者的生命体征数据构建深度学习模型,对死亡风险上升的患者发出预警,则可以节省医护人员的精力,及时对医疗设备进行合理的配置,增加医疗资源的利用率[2-7].

深度学习模型的成功应用是建立在大量带标签训练数据上的,并往往要求测试数据和训练数据服从同一分布,这在实际应用中常常不能得到满足.由于各种现实条件的限制,收集到的训练数据具有一定的局限性,例如可能某年龄段[8]、某科室或者某种并发症占据了大多数.这种局限性导致深度学习模型不能够对其他情况下的数据进行普适地预测.域适应(domain adaptation)方法能够利用源域和目的域的相似性,将源域上学习到的知识迁移到目的域上,从而解决该问题.

但是,将域适应方法应用在重症监护病人死亡风险预测任务上时还遇到主要来自3个方面的困难:整体数据分布偏移、类别之间的数据分布偏移以及时序数据的多样性和复杂性.其中整体数据分布偏移与类别之间的数据分布偏移如图1所示:

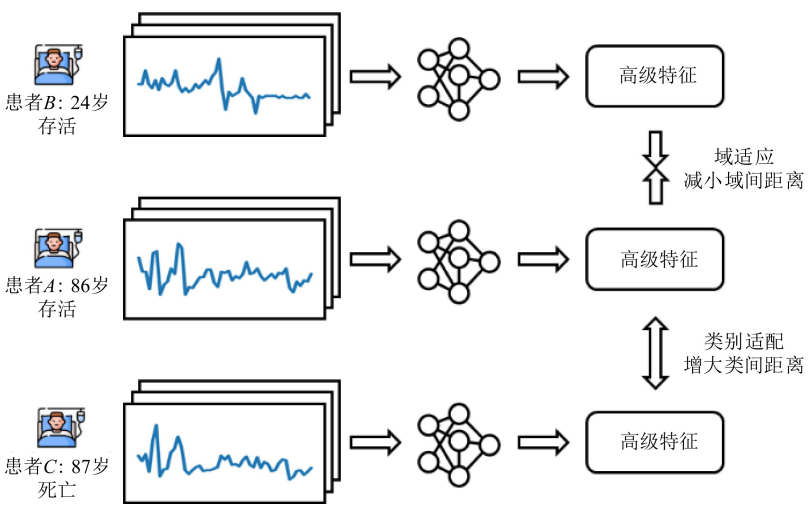

Fig. 1 Data distribution shift

图1 数据分布偏移

整体数据分布偏移指的是源域和目的域整体的数据分布往往不相同.例如,在重症监护室内收集到的数据中可能老年人占据大多数.图1中老年患者A与青年患者B的生命体征不相类似,表示以老年患者为主体的源域和以青年患者为主体的目的域的数据分布是有差异的.以医疗领域的MIMIC-Ⅲ(medical information mart for intensive care Ⅲ)数据集为例,血压作为反映患者生理状况的重要指标之一,在不同年龄段的患者之间的分布是不同的.如图2所示,患者的平均血压随着年龄增加而逐渐变低.这些生理指标分布的差异导致在老年患者数据上训练的模型不能够很好地泛化到青年患者的数据上.域适应方法能够适当减小患者A与患者B的高级特征之间的距离,消除整体数据分布偏移所带来的影响.

类别之间的数据分布偏移指的是不同域之间同一类别的数据分布往往不同.无论特征来自哪个域,相同类别的特征之间应该相隔较近,不同类别的

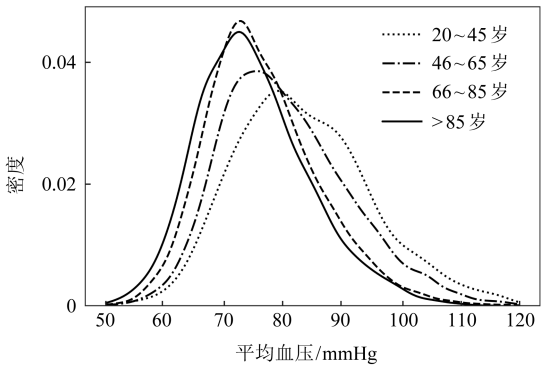

Fig. 2 Distribution of mean blood pressure with age

图 2 平均血压随年龄的分布情况

特征之间应该相隔较远.因此需要在域适应的基础上进行类别适配.如图 1 所示,患者 A 与患者 C 的年龄相近,数据分布也类似.但二者的存活结果不相同,属于不同的类,因此需要进行类别适配,增加二者高级特征之间的距离.

时序数据的多样性和复杂性所带来的困难指的是患者各项生理指标构成了数据分布互不相同的不同通道,不同通道的时序变化趋势共同描绘了病人的生理状况.深度学习模型只有在理解不同时间步之间复杂的时序依赖关系并且有效地提取高级特征之后,才能进行域适应.

本文提出了一种基于域对抗和加性余弦间隔损失的无监督域适应方法(additive margin softmax based adversarial domain adaptation, AMS-ADA).其中域对抗是一种类似生成对抗网络的方法,能够解决整体数据分布偏移的困难.加性余弦间隔损失引入了度量学习的思想,能够解决类别之间数据分布偏移带来的困难.此外,本文使用带有注意力机制的双向长短程网络作为特征提取器来应对时序数据的多样性和复杂性.

1 相关研究工作

无监督域适应问题指的是利用源域的数据和标签以及目的域的数据训练深度学习模型,希望模型能够在目的域上取得尽可能高的准确度.与许多其他的迁移学习方法[9-13]相比,域适应对目的域上的标签不做要求,进一步降低了获取标注数据的压力.深度学习模型可以简单地视作特征提取器和分类器 2 个部分.如果深度学习模型的特征提取器能够从不同域之间的数据提取出域不变(domain invariant)的特征,那么在源域上训练的分类器就可以很好地应用在目的域上.域不变的特征是指在源域和目的域都具有表现力和判别力的特征,蕴涵了源域和目的域之间可以共享的知识.为了实现提取出域不变特征的这一目标,减少整体数据分布的偏移,通常的做法有 2 种:

1) 基于特征映射的方法.对深度学习模型从源域和目的域提取出的高级特征之间施加距离约束,使得神经网络学习出的高级特征的分布相似.如 DDC (deep domain confusion)[14], DAN (deep adaptation network)[15]等方法使用了最大均值差异来衡量高级特征之间的分布差异,Deep CORAL[16]方法采用 CORAL 距离来衡量高级特征之间的分布差异.

2) 基于域对抗的方法.引入生成对抗网络的思想,用域判别器判断深度学习模型学习出的高级特征属于源域还是目的域.以对抗训练的方式使特征提取器和域判别器达到平衡.当域判别器无法辨别特征来自哪一个域的时候,说明特征提取器提取了具有域不变性的特征.如 Adversarial Discriminative Domain Adaptation[17],Domain Adversarial Neural Networks[18]等.

近年来,域对抗方法以其优异的表现而备受关注.为了减少类别之间的数据分布偏移,进一步提升无监督域适应的效果,一些工作在域对抗方法的基础上引入了度量学习的思想.例如 Wang 等人[19]和 Yin 等人[20]在域适应任务中引入了三元组损失(triplet loss),在一定程度上最小化类内距离和最大化类间距离.但是三元组损失的计算需要遍历大量样本对,增加了额外的计算量,并且需要选取合适大小的隐层特征作为三元组损失的优化对象,增加了调整超参数的负担.

2 基于域对抗和加性余弦间隔损失的无监督域适应方法

为了解决将域适应方法应用在死亡风险预测任务上时遇到的困难以及相关工作的不足,本文提出了一种基于域对抗和加性余弦间隔损失的无监督域适应方法 AMS-ADA.该方法在没有目的域样本标签的情况下,利用源域带标签的数据和目的域不带标签的数据进行训练,提升模型在目的域的准确度.该方法主要由特征提取器 G、域判别器 D 和加性余弦

间隔损失分类器 C 组成,其架构图如图 3 所示,源域和目的域数据流向分别用实线和虚线的箭头表示.

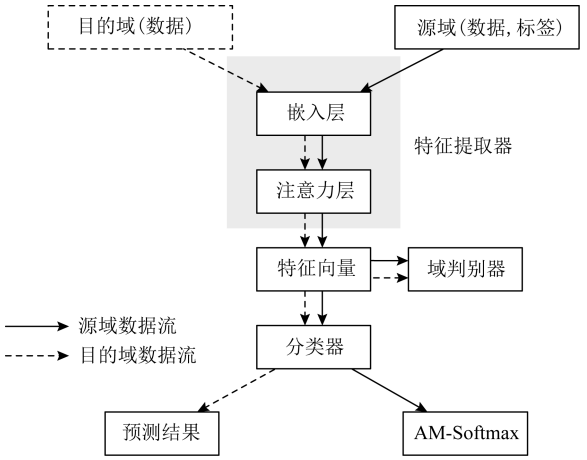

Fig. 3 Overall architecture

图 3 整体架构

2.1 问题定义

本文的研究目的是将无监督域适应方法应用在重症监护病人死亡风险预测任务上.在重症监护室内各种医疗设备每隔一段时间记录下病人的各项生命体征,这些记录可以自然地视为时序数据.

定义 1. 时序数据上的无监督域适应任务.给定有 n_s 个带标签样本的源域 $D_s=\{(\boldsymbol{x}_i^s,\boldsymbol{y}_i^s)\}_{i=0}^{n_s}$ 和 n_t 个无标签样本的目的域 $D_t=\{\boldsymbol{x}_i^t\}_{i=0}^{n_t}$,其中 $\boldsymbol{x}_i^s,\boldsymbol{x}_i^t\in\mathbb{R}^{m\times d}$ 是时序数据,d 为时序数据的长度,m 为每个时间步数据的维度.无监督域适应的目标是在 $D_s\neq D_t$ 的情况下,使目的域上的经验风险最小化.

2.2 特征提取

特征提取器负责从输入数据提取有效的高级特征.为了应对时序数据的复杂性和多样性所带来的困难,本文选取了带有注意力机制的双向长短程记忆网络作为特征提取器.其中双向长短时记忆网络(bidirectional long short term memory, BiLSTM)作为嵌入层,对输入的特征进行初步的提取,捕捉基本的时序信息.嵌入层将输入 $\boldsymbol{x}\in\mathbb{R}^{m\times d}$ 变成输出 $\boldsymbol{H}\in\mathbb{R}^{u\times d}$,即每个时间步的特征维度从 m 变为 u,并且包含了上下文的信息.

为了更好地提取时序信息,本文使用了自注意力机制[21],对嵌入层输出的每一个时间步计算注意力值 $a_i,i=1,2,\cdots,d$,再根据注意力值对所有时间步进行加权求和.注意力机制能够使得深度学习模型更关注重要的时间步,从而能够提取出表现力更强的特征.

记 $\boldsymbol{W}_1\in\mathbb{R}^{n_a\times u},\boldsymbol{W}_2\in\mathbb{R}^{r\times n_a}$ 为参数矩阵,n_a 为计算注意力的隐层向量维度,r 为注意力头的个数.Softmax 操作对每个行向量进行,目的是使得每个时间步的注意力值的和为 1.注意力矩阵 $\boldsymbol{A}\in\mathbb{R}^{r\times d}$ 的计算方式表示为

$$\boldsymbol{A}=\mathrm{Softmax}(\boldsymbol{W}_2\tanh(\boldsymbol{W}_1\boldsymbol{H})). \tag{1}$$

最后,注意力层的输出即为整个特征提取器的输出 $G(\boldsymbol{x})\in\mathbb{R}^{r\times u}$,表示为

$$\boldsymbol{M}=\boldsymbol{A}\boldsymbol{H}^{\mathrm{T}}. \tag{2}$$

2.3 域对抗

域判别器的作用是以域对抗的形式进行域适应,学习到域不变的特征,试图解决整体数据分布偏移的问题.域对抗借鉴了生成对抗网络的思想,使特征提取器和域判别器之间相互竞争,当域判别器无法辨别特征来自源于还是目的域时,特征提取器学会了如何提取域不变的特征.

记源域和目的域的概率分布为 $p(X_s)$ 和 $p(X_t)$,域判别器 D 的优化目标可以表达为

$$\begin{aligned}\max_D V(D)=&E_{\boldsymbol{x}^t\sim p(X_t)}[\log D(G(\boldsymbol{x}^t))]+\\&E_{\boldsymbol{x}^s\sim p(X_s)}[\log(1-D(G(\boldsymbol{x}^s)))].\end{aligned} \tag{3}$$

特征提取器 G 的优化目标可表示为

$$\begin{aligned}\min_G V(G)=&E_{\boldsymbol{x}^t\sim p(X_t)}[\log D(G(\boldsymbol{x}^t))]+\\&E_{\boldsymbol{x}^s\sim p(X_s)}[\log(1-D(G(\boldsymbol{x}^s)))].\end{aligned} \tag{4}$$

特征提取器和域判别器的优化目标可以结合在一起,写成极小极大的优化形式:

$$\begin{aligned}\min_G\max_D V(D,G)=&E_{\boldsymbol{x}^t\sim p(X_t)}[\log D(G(\boldsymbol{x}^t))]+\\&E_{\boldsymbol{x}^s\sim p(X_s)}[\log(1-D(G(\boldsymbol{x}^s)))].\end{aligned} \tag{5}$$

特征提取器和域判别器都是深度学习模型,在实践中通常以梯度下降最小化损失函数的形式进行优化.记特征提取器和域判别器的模型参数为 θ_G 和 θ_D,域判别器的损失函数 $L_{\mathrm{disc}}(\theta_G,\theta_D)$ 可以写为

$$\begin{aligned}L_{\mathrm{disc}}(\theta_G,\theta_D)=&-\frac{1}{n_s}\sum_{i=1}^{n_s}\log(1-D(G(\boldsymbol{x}^s)))-\\&\frac{1}{n_t}\sum_{i=1}^{n_t}\log D(G(\boldsymbol{x}^t)).\end{aligned} \tag{6}$$

在对抗训练的过程中,特征提取器和域判别器的优化是交替进行的,形式化地表达为

$$\begin{aligned}\hat{\theta}_D&=\arg\min_{\theta_D}L_{\mathrm{disc}}(\theta_G,\theta_D),\\\hat{\theta}_G&=\arg\max_{\theta_G}L_{\mathrm{disc}}(\theta_G,\theta_D).\end{aligned} \tag{7}$$

以域对抗的方式进行域适应,能够利用生成对抗网络强大的拟合数据分布的能力,更好地提取出域不变的特征.

2.4 加性余弦间隔损失分类器

加性余弦间隔(additive margin softmax, AM-Softmax)损失引入了度量学习的思想,能够增强不同类别的样本之间的可区分性.它作为最终的分类损失函数,能够同时端到端地最小化类内距离和最大化类间距离,不需要再耗费时间去选取深度学习模型中哪一层的特征作为优化目标.相比于三元组损失函数,它不需要额外计算样本对之间的距离,节省了训练所需时间.此外,在角度空间端到端地对类内距离和类间距离进行优化相比于三元组损失对隐层向量进行优化能取得更好的效果.接下来以对Softmax损失进行改进的形式介绍加性余弦间隔损失的动机和原理.

记 n 为当前批次训练样本的数量,y_i 为样本 $\boldsymbol{x}_i$ 的类别标签,共有 c 类.$\mathbb{I}(\cdot)$为示性函数,当括号内表达式为真时其值为1,当表达式为假时其值为0.$p(j\mid\boldsymbol{x}_i)$为模型给出的样本 $\boldsymbol{x}_i$ 属于第 j 类的概率.Softmax损失函数 L_{S} 可以写为

$$L_{\mathrm{S}}=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mathbb{I}(y_i=j)\log p(j\mid\boldsymbol{x}_i). \tag{8}$$

记样本 $\boldsymbol{x}_i$ 对在深度学习模型中最后一层的输入为 $\boldsymbol{f}_i$,$\boldsymbol{W}$ 为最后一层的权重矩阵,$\boldsymbol{W}_j$ 为权重矩阵中对应输出类别 j 的行向量.省略偏置项,Softmax损失函数进一步写为

$$L_{\mathrm{S}}=-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c}\mathbb{I}(y_i=j)\log\frac{\mathrm{e}^{\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{f}_i}}{\sum_{k=1}^{c}\mathrm{e}^{\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{f}_i}}=$$
$$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i}}{\sum_{j=1}^{c}\mathrm{e}^{\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{f}_i}}. \tag{9}$$

记 $\boldsymbol{f}_i$ 与 $\boldsymbol{W}_j$ 的夹角为 $\cos\theta_{i,j}$,对权重矩阵和输入进行归一化,即令 $\|\boldsymbol{f}_i\|=1$,$\|\boldsymbol{W}_j\|=1$.记缩放值为 η,Softmax函数可以用余弦值来表示:

$$L_{\mathrm{S}}=-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{f}_i\|\cos\theta_{i,y_i}}}{\sum_{j=1}^{c}\mathrm{e}^{\|\boldsymbol{W}_j\|\|\boldsymbol{f}_i\|\cos\theta_{i,j}}}=$$
$$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\eta\cos\theta_{i,y_i}}}{\sum_{j=1}^{c}\mathrm{e}^{\eta\cos\theta_{i,j}}}. \tag{10}$$

将向量内积写成夹角的形式,使得对决策边界的分析从欧氏空间转变为角度空间.现在以二分类的场景对决策边界进行分析,如图4所示.此时类别数 $c=2$.当 $\cos\theta_{i,0}>\cos\theta_{i,1}$ 时,样本 $\boldsymbol{x}_i$ 被判定为 c_0 类.同理,当 $\cos\theta_{i,1}>\cos\theta_{i,0}$ 时,样本 $\boldsymbol{x}_i$ 被判定为 c_1 类.当前情况下,Softmax损失能够为不同类别划分清晰的界限,但是没有显式地优化类间的离散度度以及类内的聚合度.为了增加决策边界的宽度,引入边界阈值 m.现在对决策边界施加更加严格的要求,当 $\cos\theta_{i,0}-m>\cos\theta_{i,1}$ 时,样本 $\boldsymbol{x}_i$ 被判定为 c_0 类,当 $\cos\theta_{i,1}-m>\cos\theta_{i,0}$ 时,样本 $\boldsymbol{x}_i$ 被判定为 c_1 类.将二分类的情况推广为多分类便可得到加性余弦间隔损失.记特征提取器和分类器的参数分别为 θ_G 和 θ_C,加性余弦间隔损失 $L_{\mathrm{AMS}}(\theta_G,\theta_C)$ 形式化地表达为

$$L_{\mathrm{AMS}}(\theta_G,\theta_C)=$$
$$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\eta(\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i-m)}}{\mathrm{e}^{\eta(\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{f}_i-m)}+\sum_{j=1,j\neq y_i}^{c}\mathrm{e}^{\eta\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{f}_i}}=$$
$$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{e}^{\eta(\cos\theta_{i,y_i}-m)}}{\mathrm{e}^{\eta(\cos\theta_{i,y_i}-m)}+\sum_{j=1,j\neq y_i}^{c}\mathrm{e}^{\eta\cos\theta_{i,j}}}. \tag{11}$$

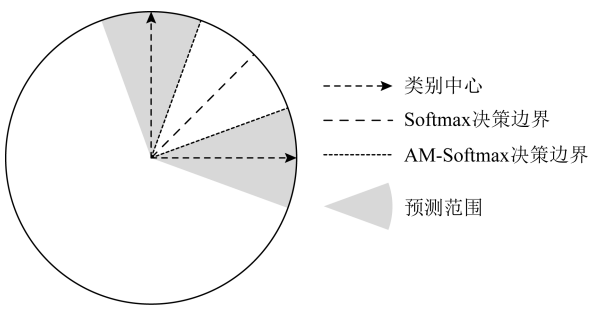

Fig. 4 Comparison between AM-Softmax Loss and Softmax Loss

图4 加性余弦间隔损失和Softmax损失的对比

对决策边界施加的限制能够在角度空间最大化分类器的决策边界,从而达到最小化类内距离和最大化类间距离的目的.

2.5 训练流程

本文提出的方法含有可训练参数的部分为特征提取器、域判别器和分类器,其参数分别记为 θ_G,θ_D,θ_C.由式(6)和式(11)可得最终的损失函数 $L(\theta_G,\theta_C,\theta_D)$:

$$L(\theta_G,\theta_C,\theta_D)=L_{\mathrm{AMS}}(\theta_G,\theta_C)-\lambda L_{\mathrm{disc}}(\theta_G,\theta_D), \tag{12}$$

其中 λ 为平衡因子,调节 L_{AMS} 和 L_{disc} 的比例.

本文提出方法的详细训练流程如AMS-ADA算法所示.首先对特征提取器、域判别器和分类器的参数进行随机的初始化.训练过程中对这些参数以梯度下降的形式进行交替优化.本文选用深度学习领域中常用的 *Adam* 优化器完成梯度下降的任务.在对抗训练的每次迭代的过程中,为了使得域判别器能够更好地指导特征提取器生成域不变的特征,

需要增加域判别器的更新次数,即域判别器更新 N_{disc} 次之后,特征提取器和分类器才更新一次.域判别器的更新是指计算 $L_{\text{disc}}(\theta_G,\theta_D)$ 后通过反向传播更新域判别器的参数.特征提取器和分类器的更新也是类似地计算各自的损失函数后通过反向传播对参数进行更新.当损失函数收敛之后,得到训练好的模型.此时将目的域的数据输入模型,得到最终的预测值.

算法 1. 基于域对抗和加性余弦间隔损失的无监督域适应方法.

输入:带标签的源域数据 $D_{\text{s}}=\{(x_i^{\text{s}},y_i^{\text{s}})\}_{i=0}^{n_{\text{s}}}$,无标签的目的域数据 $D_{\text{t}}=\{x_i^{\text{t}}\}_{i=0}^{n_{\text{t}}}$,平衡因子 λ,特征提取器、域判别器和分类器的参数 $\theta_G,\theta_D,\theta_C$,每次迭代中域判别器的更新次数 N_{disc},用于梯度下降的优化器 $Adam$;

输出:对目的域数据的预测值 $\{y_i^{\text{t}}\}_{i=0}^{n_{\text{t}}}$,优化后的特征提取器、域判别器和分类器的参数 $\hat{\theta}_G,\hat{\theta}_D,\hat{\theta}_C$.

① 随机初始化 $\theta_G,\theta_D,\theta_C$;

② repeat

③ for $i=1,2,\cdots,N_{\text{disc}}$ do

④ 根据式(6)计算 $L_{\text{disc}}(\theta_G,\theta_D)$;

⑤ $\theta_D\leftarrow Adam(\nabla_{\theta_D}L_{\text{disc}}(\theta_G,\theta_D))$;

⑥ end for

⑦ 根据式(12)计算 $L(\theta_G,\theta_C,\theta_D)$;

⑧ $\theta_C\leftarrow Adam(\nabla_{\theta_C}L(\theta_G,\theta_C,\theta_D))$;

⑨ $\theta_G\leftarrow Adam(\nabla_{\theta_G}L(\theta_G,\theta_C,\theta_D))$;

⑩ until 模型参数收敛

⑪ $\hat{\theta}_G,\hat{\theta}_D,\hat{\theta}_C\leftarrow\theta_G,\theta_D,\theta_C$;

⑫ $\{y_i^{\text{t}}\}_{i=0}^{n_{\text{t}}}\leftarrow\{C(G(x_i^{\text{t}}))\}_{i=0}^{n_{\text{t}}}$;

⑬ return $\{y_i^{\text{t}}\}_{i=0}^{n_{\text{t}}},\hat{\theta}_G,\hat{\theta}_D,\hat{\theta}_C$.

3 实 验

本文选用 MIMIC-Ⅲ数据集[22]进行实验.MIMIC-Ⅲ数据集是麻省理工大学维护的公共临床数据库,包含 2001—2016 年之间约 6 万例的住院记录,每条记录包括人口统计特征、医疗干预记录、成像报告、生命体征记录、护理记录等信息.

Harutyunyan 等人[2]在 MIMIC-Ⅲ数据集的基础上定义了死亡风险预测任务.一般来说,患者进入重症监护室后的 48 h 以内的情况较为危急,因此本文选取患者进入重症监护室之后的 48 h 以内的数据对患者的存活结果进行预测.

根据 Harutyunyan 等人[2]的工作,本文在 MIMIC-Ⅲ数据集中提取了 76 维的特征,包括心率(heart rate)、舒张压(systolic blood pressure)、收缩压(diastolic blood pressure)、血氧饱和度(SpO2)、毛细血管填充率(capillary refill rate)等 60 维的连续特征和格拉斯哥昏迷指数(Glasgow coma scale)等 12 维的离散特征以及 4 维的关于患者信息的常量.经过数据清洗和预处理后,最终得到的输入数据共有 48 个时间步,每个时间步有 76 维的特征.

Purushotham 等人[23]尝试在不同年龄段的急性低氧性呼吸衰竭患者之间进行了迁移学习.本文沿用了该文的实验设置,将 MIMIC-Ⅲ数据集的 ICU 数据库中所有患者按照年龄分为 4 组,如表 1 所示:

Table 1 Different Domains of MIMIC-Ⅲ Dataset

表 1 MIMIC-Ⅲ数据集不同域的划分

域名称	年龄段	人员数量
青年	20～45	2 763
中年	46～65	6 963
老年	66～85	9 100
高龄老年	>85	2 313

由于数据集中的正负样本比例相差较大,且属于二分类问题,为了避免正负样本不均衡对评价指标带来的影响,本次实验采用 ROC 曲线(receiver operating characteristic curve)下的面积值(area under curve, AUC)作为评价标准.本文采用了 5 种方法与本文提出的 AMS-ADA 方法进行对比:

1) BiLSTM.使用结合自注意力机制的 BiLSTM 网络作为基线,在源域上训练,在目的域上测试,没有使用任何无监督域适应学习方法.

2) CORAL.使用结合自注意力机制的 BiLSTM 网络提取特征,采用基于特征映射的迁移学习方法 Deep CORAL[16].该方法以 CORAL 距离衡量源域特征分布和目的域特征分布的差异.

3) DAN.使用结合自注意力机制的 BiLSTM 网络提取特征,采用基于特征映射的迁移学习方法 DAN[15].该方法以最大均值差异衡量源域特征分布和目的域特征分布的差异.

4) ADA(adversarial domain adaptation).与 1)～3)所述方法采用相同的特征提取器,并且使用了域对抗方法,使用 Softmax 损失函数.

5) Tri-ADA(triplet loss guided adversarial

domain adaptation)[19].与 1)～4)所述方法使用相同的特征提取器.使用了域对抗方法.并且在此基础上加上三元组损失函数,以解决类别之间的数据分布偏移的问题.

本文在 4 个不同年龄段,即 4 个域之间两两进行无监督域适应任务,实验结果如表 2 所示.例如将青年患者的数据作为源域,将中年患者的数据作为目的域,无监督域适应任务记为青年→中年.

Table 2 Experimental Results of Mortality Prediction Task Based on Unsupervised Domain Adaptation

表 2 基于无监督域适应的死亡风险预测实验结果

源域→目的域	BiLSTM	CORAL	DAN	ADA	Tri-ADA	AMS-ADA
青年→中年	0.803	0.822	0.832	0.834	0.831	**0.842**
青年→老年	0.781	0.794	0.781	0.790	0.795	**0.808**
青年→高龄老年	0.756	0.763	0.759	0.766	0.765	**0.781**
中年→青年	0.867	0.874	0.879	0.883	0.880	**0.889**
中年→老年	0.808	0.825	0.828	0.831	**0.835**	**0.835**
中年→高龄老年	0.754	0.766	0.779	0.784	0.780	**0.793**
老年→青年	0.845	0.863	0.870	**0.875**	0.873	0.868
老年→中年	0.832	0.848	0.853	0.857	0.858	**0.862**
老年→高龄老年	0.763	0.792	0.794	0.797	0.795	**0.807**
高龄老年→青年	0.826	0.841	0.840	0.856	**0.861**	0.856
高龄老年→中年	0.804	0.801	0.796	0.799	0.812	**0.819**
高龄老年→老年	0.775	0.802	0.799	0.805	0.803	**0.815**

注:使用的评价标准为 AUC 值,黑体数据表示最佳结果.

本文提出的 AMS-ADA 方法在 12 个无监督域适应任务中的 10 个取得了最高的 AUC 值,说明了该方法的有效性.BiLSTM 方法没有使用任何无监督域适应方法,因此表现较差.对于相隔较远的域,BiLSTM 方法的表现下降较为明显.比如对于任务中年→青年,BiLSTM 方法的 AUC 值为 0.867,而对于任务中年→高龄老年,BiLSTM 方法的 AUC 值下降为 0.754.相隔较远的域意味着年龄相隔较大,数据的分布差异更为显著,因此对模型的准确度影响较大.CORAL 方法和 DAN 方法使用了基于特征映射的迁移方法尝试解决全局的数据分布差异的问题,从结果上可以看出这 2 种方法相比 BiLSTM 方法有一定的提升.ADA 方法引入了域对抗,相比于基于特征映射的方法能够更好地减少全局的数据分布差异,因此效果更好.Tri-ADA 以域对抗的形式进行域适应,并且加入了三元组损失以减少类别之间的数据分布差异.实验结果较之 CORAL 和 DAN 方法有了一定的提升.为了更精细地对齐类别之间的数据分布,本文提出的 AMS-ADA 方法引入了加性余弦间隔损失,相比 ADA 方法和 Tri-ADA 方法的准确度有了进一步的提升,说明了本文提出方法的有效性.

为了直观体现本文提出方法的优越性,分别训练 BiLSTM,Tri-ADA,AMS-ADA 这 3 种方法,取各个方法的分类器的最后一层输出特征投影到角度空间进行可视化.选取青年患者的数据作为源域,高龄老年患者的数据作为目的域.BiLSTM 方法的源域和目的域特征可视化结果分别如图 5 和图 6 所示.BiLSTM 方法在目的域的准确度下降,其原因之一是类别之间的分布偏移.在源域训练时,不同类别的特征之间具有明显的界限.但是决策边界不够宽,在目的域测试时由于分布偏移导致分类错误.

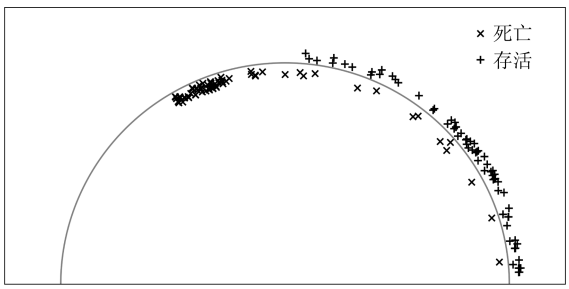

Fig. 5 Source domain feature visualization of BiLSTM method

图 5 BiLSTM 方法的源域特征可视化

因此,模型应该显式地增大决策边界,保持类内紧凑性和类间可分离性.Tri-ADA 方法的源域和目的域特征可视化结果分别如图 7 和图 8 所示.Tri-

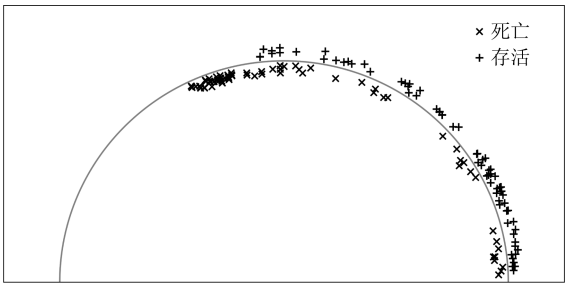

Fig. 6 Target domain feature visualization of BiLSTM method

图 6 BiLSTM 方法的目的域特征可视化

ADA 方法在源域训练时以三元组损失的形式增大了类间距离,因此在目的域测试时不同类别的特征之间可分离性加强,从而降低了错误率.

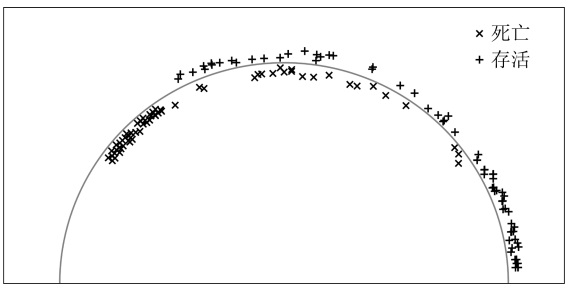

Fig. 7 Source domain feature visualization of Tri-ADA method

图 7 Tri-ADA 方法的源域特征可视化

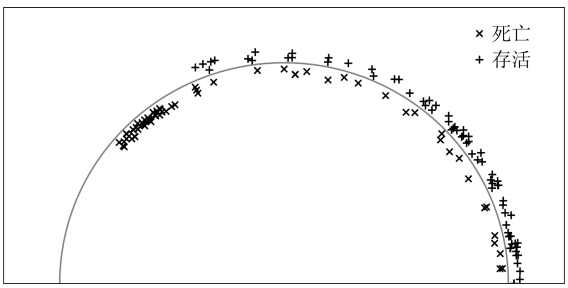

Fig. 8 Target domain feature visualization of Tri-ADA method

图 8 Tri-ADA 方法的目的域特征可视化

AMS-ADA 方法引入了 AM-Softmax 损失函数,能够进一步在角度空间增加决策边界的宽度,其源域和目的域特征可视化结果分别如图 9 和图 10 所示.存活患者的特征与死亡患者的特征的重叠部分进一步缩小,取得了很好的类间可分离性和类内紧凑性.得益于更宽的决策边界,在源域上训练的分类器对类别偏移的敏感程度下降,因此在目的域上测试时能够取得更好的准确度.

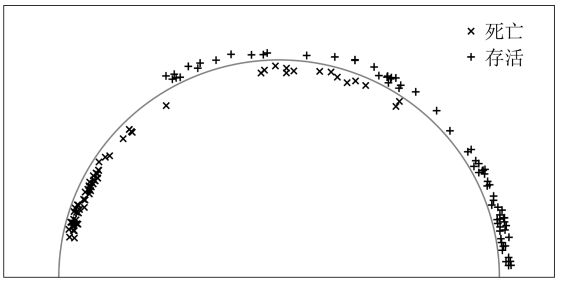

Fig. 9 Source domain feature visualization of AMS-ADA method

图 9 AMS-ADA 方法的源域特征可视化

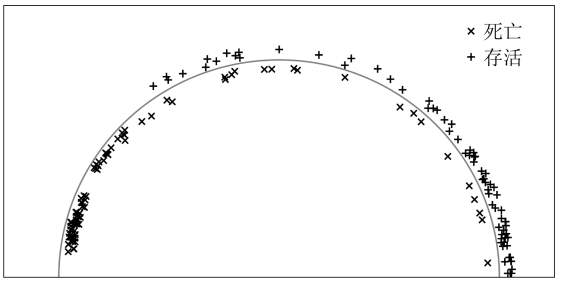

Fig. 10 Target domain feature visualization of AMS-ADA method

图 10 AMS-ADA 方法的目的域特征可视化

4 结束语

深度学习模型的实际应用中容易遇到训练数据不足、整体数据分布偏移和类别之间数据分布偏移的问题.本文提出了一种基于域对抗和加性余弦间隔损失的无监督域适应方法应对这些问题.本文以域对抗的形式减少了整体数据之间数据分布偏移.为了进一步改善无监督域适应的效果,引入度量学习的思想,以最小化加性余弦间隔损失的形式减少了类别之间的数据分布偏移.所提出的方法在重症监护病人死亡风险预测任务上进行了验证,在 MIMIC-Ⅲ数据集上的实验结果和可视化分析结果证明了该方法的有效性.未来的工作会尝试将所提出方法扩展到医疗领域的其他任务中,例如疾病预测和住院时长预测等任务.

作者贡献声明:蔡德润提出算法思路、完成实验并撰写论文;李红燕提出了指导意见并修改论文.

参 考 文 献

[1] Huang Chaolin, Wang Yeming, Li Xingwang, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China [J]. The Lancet, 2020, 395(10223): 497-506

[2] Harutyunyan H, Khachatrian H, Kale D C, et al. Multitask learning and benchmarking with clinical time series data [J]. Scientific Data, 2019, 6(1): 1-18

[3] Hong Shenda, Xu Yanbo, Khare A, et al. HOLMES: Health online model ensemble serving for deep learning models in intensive care units [C] //Proc of the 26th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2020: 1614-1624

[4] Lipton Z C, Kale D C, Elkan C, et al. Learning to diagnose with LSTM recurrent neural networks [C/OL] //Proc of 2016 Int Conf on Learning Representations(ICLR). (2017-03-21) [2020-12-25]. https://arxiv.org/abs/1511.03677

[5] Che Zhengping, Kale D, Li Wenzhe, et al. Deep computational phenotyping [C] //Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2015: 507-516

[6] Liu Zitao, Hauskrecht M. Learning adaptive forecasting models from irregularly sampled multivariate clinical data [C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 1273-1279

[7] Moor M, Horn M, Rieck B, et al. Early recognition of sepsis with Gaussian process temporal convolutional networks and dynamic time warping [C] //Proc of the 4th Machine Learning for Healthcare Conf. Cambridge, MA: JMLR, 2019: 2-26

[8] Alemayehu B, Warner K E. The lifetime distribution of health care costs [J]. Health Services Research, 2004, 39(3): 627-642

[9] Xiao Cao, Nghia H T, Hong Shenda, et al. CHEER: Rich model helps poor model via knowledge infusion[J/OL]. IEEE Transactions on Knowledge and Data Engineering, 2020 [2020-12-25]. http://dx.doi.org/10.1109/TKDE.2020.2989405

[10] Hong Shenda, Xiao Cao, Hoang T N, et al. RDPD: Rich data helps poor data via imitation [C] //Proc of the 28th Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 5895-5901

[11] Gupta P, Malhotra P, Narwariya J, et al. Transfer learning for clinical time series analysis using deep neural networks [J]. Journal of Healthcare Informatics Research, 2020, 4(2): 112-137

[12] Gupta P, Malhotra P, Vig L, et al. Using features from pre-trained TimeNET for clinical predictions [C] //Proc of the 3rd Int Workshop on Knowledge Discovery in Healthcare Data at IJCAI. Menlo Park, CA: AAAI, 2018: 38-44

[13] Islam K A, Hill V, Schaeffer B, et al. Semi-supervised adversarial domain adaptation for seagrass detection using multispectral images in coastal areas [J]. Data Science and Engineering, 2020, 5: 111-125

[14] Tzeng E, Hoffman J, Zhang Ning, et al. Deep domain confusion: Maximizing for domain invariance [J]. arXiv preprint, arXiv:1412.3474, 2014

[15] Long Mingsheng, Cao Yue, Wang Jianmin, et al. Learning transferable features with deep adaptation networks [C] // Proc of the 32nd Int Conf on on Machine Learning. Cambridge, MA: JMLR, 2015: 97-105

[16] Sun Baochen, Saenko K. Deep CORAL: Correlation alignment for deep domain adaptation [C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 443-450

[17] Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7167-7176

[18] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks [J]. Journal of Machine Learning Research, 2016, 17(1): 2096-2030

[19] Wang Xiaodong, Liu Feng. Triplet loss guided adversarial domain adaptation for bearing fault diagnosis [J]. Sensors, 2020, 20(1): 320-339

[20] Yin Yueming, Yang Zhen, Hu Haifeng, et al. Metric-learning-assisted domain adaptation [J]. arXiv preprint, arXiv:2004.10963, 2020

[21] Lin Zhouhan, Feng Minwei, Santos C N, et al. A structured self-attentive sentence embedding [C/OL] //Proc of 2017 Int Conf on Learning Representations (ICLR). (2017-03-09) [2020-12-25]. https://arxiv.org/abs/1703.03130

[22] Johnson A E W, Pollard T J, Shen Lu, et al. MIMIC-Ⅲ, a freely accessible critical care database [J]. Scientific Data, 2016, 3(1): 1-9

[23] Purushotham S, Carvalho W, Nilanon T, et al. Variational recurrent adversarial deep domain adaptation [C/OL] //Proc of 2017 Int Conf on Learning Representations(ICLR). [2020-12-25]. https://openreview.net/pdf?id=rk9eAFcxg

Cai Derun, born in 1998. Master. His main research interests include data mining and knowledge discovery.

蔡德润,1998 年生.硕士.主要研究方向为数据挖掘与知识发现.

Li Hongyan, born in 1970. PhD, professor, PhD supervisor. Her main research interests include data management, data mining and knowledge discovery.

李红燕,1970 年生.博士,教授,博士生导师.主要研究方向为数据管理、数据挖掘与知识发现.