计算机科学
COMPUTER SCIENCE

针对多标记表格数据的半监督学习方法

葛泽庆, 黄圣君

引用本文

葛泽庆, 黄圣君. 针对多标记表格数据的半监督学习方法[J]. 计算机科学, 2026, 53(3): 151-157.

GE Zeqing, HUANG Shengjun. Semi-supervised Learning Method for Multi-label Tabular Data[J]. Computer Science, 2026, 53(3): 151-157.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250600149

针对多标记表格数据的半监督学习方法

葛泽庆　黄圣君
南京航空航天大学计算机科学与技术学院　南京 211106
(gezeqing@nuaa.edu.cn)

摘　要　表格数据在医学、金融和制造业等领域具有广泛应用,其多标记分类任务对揭示现实世界中复杂的关联特性至关重要。然而,获取大规模标记数据集往往成本高昂,这给研究带来了挑战。虽然半监督学习利用未标记样本在图像和文本数据中取得了成功,但由于表格数据缺乏固有的空间或语义结构,使得传统方法效率较低。为了应对这些挑战,提出了一种针对多标记表格数据的半监督学习框架。该方法引入了一种结构保留的数据增强方法,在特征表示空间内添加高斯噪声保留原始数据结构,与基于一致性的正则化技术,在样本及其扰动版本之间进行正则化,以增强泛化能力。此外,还开发了一种基于注意力机制的机制,有选择地从标记数据中聚合邻域信息,从而使模型能够有效地利用局部特征相关性。在10个公共多标记表格数据集上进行了广泛的实验,结果证明了该方法的有效性。

关键词:表格数据;多标记分类;半监督学习;数据增强;注意力机制
中图分类号　TP181

Semi-supervised Learning Method for Multi-label Tabular Data

GE Zeqing and HUANG Shengjun
College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China

Abstract　Tabular data is ubiquitous in industrial applications, spanning fields such as medicine, finance, and manufacturing, where each sample is characterized by heterogeneous features. Multi-label classification for tabular data is crucial for capturing the complex, interconnected nature of real-world phenomena, yet obtaining large-scale labeled datasets is often costly. While semi-supervised learning has shown success in image and text data by leveraging unlabeled samples, its application to tabular data remains challenging due to the lack of inherent spatial or semantic structures, making conventional augmentation and consistency-based methods less effective. To address these challenges, this paper proposes a novel semi-supervised learning framework tailored for multi-label tabular data. This approach introduces a structure-preserving data augmentation method that adds Gaussian noise to the feature representation space preserving the original data structure, and a consistency-based regularization technique between samples and their perturbed versions to enhance generalization. Additionally, an attention-based mechanism is developed to selectively aggregate neighborhood information from labeled data, allowing the model to leverage local feature correlations effectively. For unlabeled data, a state-of-the-art pseudo-labeling strategy is employed to enable iterative refinement of model predictions. Extensive experiments are conducted on ten public multi-label tabular datasets, covering various domains to validate the robustness of the proposed method. Results demonstrate the effectiveness of the proposed method, advancing the state of semi-supervised multi-label learning for tabular data.

Keywords　Tabular data, Multi-label classification, Semi-supervised learning, Data augmentation, Attention mechanism

1　引言

机器学习(ML)问题中,表格数据是一种常见的数据类型,广泛应用于各种工业应用,包括医学、金融、制造业和其他领域[1]。在这种类型的数据中,每个对象都由一组异构特征描述。这些特征可以是数值、分类和时间序列数据的混合,反映了对象的多维属性。在医学领域,表格数据可能包括患者的年龄、性别、血压和药物使用情况等信息。在金融领域,它可能包括客户交易历史、信用评分和收入水平的数据。在制造业,它可能涉及生产日期、材料类型和质量检验结果等细节。这些多样化的特征为机器学习模型提供了丰富的信息来源,使它们能够执行复杂的任务,例如回归、分类和优化。鉴

到稿日期:2025-06-24　返修日期:2025-08-22
基金项目:国家自然科学基金优秀青年科学基金(62222605);叶企孙基金(U2441285)
Thls work was supported by the Excellent Young Scientists Found of the National Natural Science Foundation of China(62222605) and YQS Foundation(U2441285).
通信作者:黄圣君(huangsj@nuaa.edu.cn)

于表格数据在这些关键领域无处不在，开发和应用有效的机器学习算法来处理和分析表格数据变得越来越重要。

在实际应用中，单个样本对应多个标记的任务对于图像和文本数据越来越常见且重要，对于表格数据也是如此[2]。例如，在一张照片中，可能有一只狗、一只猫和一棵树，每个都需要单独识别；而一篇关于气候变化的文章可能同时涵盖环境科学、政策和经济学。同样，在表格数据中，单个记录可以与多个标记相关联，同时反映数据的各个方面。例如，在医疗保健领域，患者的记录可能表明多种病症、治疗方法和结果；在金融领域，交易可能涉及多种风险因素、投资类型和合规性检查。在这些情况下，多标记分析提供了更全面的理解和更优的决策能力。

以表格形式处理多标记数据的能力至关重要，因为它反映了现实世界现象的复杂性和相互关联性。通过利用多标记分类，机器学习模型可以提供更细致入微、更可操作的见解，推动医疗保健、金融、制造业等各个领域的进步。这使得多标记任务不仅更加普遍，而且在实际应用中也更加实用和有影响力。

然而，多标记分类任务的训练需要大量标记数据，这可能导致注释成本过高。这一挑战催生了半监督多标记学习。在图像和文本数据的背景下，半监督学习已成功利用未标记数据来提高模型性能[3]。例如，伪标记[4]、数据增强[5]和一致性正则化[6]等技术已显现出巨大的前景。图像数据可以从翻转、裁剪或颜色抖动等增强中受益[7]，而文本数据可以利用同义词替换或反向翻译进行增强[8]。这些方法通过从现有数据中创建不同的训练实例，帮助模型更好地学习。然而，与具有空间信息的图像数据或具有语义信息的自然语言数据不同，表格数据缺乏这种明显的特征关系。这种固有的复杂性使得难以在表格数据中有效应用传统的数据增强技术，与图像或文本不同，表格数据同时包含数值与类别特征，其结构对传统的随机转换增强技术并不友好，直接应用可能会破坏表格数据的完整性。此外，在半监督场景下有效利用大量未标记样本面临相当大的挑战，需要深入理解表格数据结构，设计复杂的算法准确地从未标记数据中推断出有效的信息，并将其与已标注数据集结合以提高模型性能。直接应用传统方法，如暂时集成(Temporal Ensembling)[9]，将每个示例的预测的指数滑动平均(EMA)保留为教师模型，可能无法为表格数据带来改进。

为了应用上述挑战，本文设计了新的数据增强方法，该方法不会破坏原始数据结构，且在样本及其噪声版本之间进行基于一致性的正则化，以实现在多标记表格数据上进行半监督学习，提高模型的泛化性能。此外，本文还开发了一种算法，通过注意力机制，利用来自标记数据的近邻信息来提高模型的预测精度。对于未标记数据，使用先进的伪标记策略来继续使用监督损失。在 10 个公共数据集上的大量实验结果有效地验证了本文方法对于表格半监督多标记学习的有效性。

2 相关工作

本文方法针对半监督场景下的表格多标记数据，相关工作介绍围绕表格数据的已有数据增强方法、表格数据的多标记方法以及半监督场景下的多标记方法这 3 点展开。

2.1 表格数据的数据增强

表格数据的数据增强主要被提出作为自监督学习方法的辅助技术。本文的重点仅放在数据增强技术上，而不深入研究后续的自监督方法。例如，VIME[10]和 SCARF[11]引入了一种数据增强技术，该技术随机掩码输入数据并通过从同一批样本中采样替换掩码的值。SAINT[12]使用 cutmix 和 manifold mixup 来创建数据的不同视图。ExcelFormer[13]与 mixup 相比考虑到了特征重要性，其在输入层引入了 FeatMix，在隐藏层引入了 HiddenMix，在两个样本之间随机交换一些嵌入特征。但是，上述方法是基于插值的，同时修改了样本标记。

2.2 表格数据多标记学习

早期的工作并没有训练模型，而是通过惰性学习来完成多标记分类任务。最早的多标记惰性学习方法是 ML-KNN[14]，它使用最大后验原理来识别未见过的样本的标记。随着深度学习的发展，处理多标记数据最直接的策略是在推导分类模型中利用实例的相同表示，但是由于其忽视了每个类别标记独特的性质，可能不能达到最优。DELA[15]从对偶角度应对特定标记特征学习的挑战，它尝试识别每个标记独有的非信息性特征，然后使分类器对这些识别的特征具有不变性。

2.3 半监督多标记学习

半监督多标记学习通常可以使用标记数据来训练模型，给未标记数据打上伪标记，以便使用监督学习框架进行进一步训练。COMN[16]使用一对具有两组不同参数的 ML-KNN 分类器在同一数据集上进行训练。两个分类器都对未标记实例进行标注并相互提供训练数据集。随后，大量工作致力于提高伪标记的质量。CAP[17]引入了一个包含类别感知阈值的正则化学习框架，可有效控制每个类别的正负伪标记分配。PCLP[18]利用基于结构因果模型推断的相关性诱导标签先验，约束和引导伪标签的生成，提高了伪标记的可靠性。

3 针对多标记表格数据的半监督学习方法

3.1 预备知识

令 $\mathcal{X}=R^d$ 表示 d 维输入空间，$\mathcal{Y}=\{0,1\}^t$ 表示 t 个可能类别的标记输出空间，则多标记示例表示为 $(\boldsymbol{x},\boldsymbol{y})$。其中 $\boldsymbol{x}\in\mathcal{X}$ 是其特征向量，$y\in\mathcal{Y}$ 是其相关标记。$\boldsymbol{y}=[y_1,y_2,\cdots,y_t]\in\{0,1\}^t$ 是一个 t 维向量，其中 $y_k=1$ 表示第 i 个标记与该实例相关，否则 $y_k=0$。在半监督学习场景中，训练集包含 n 个有标记的训练样例 $\mathcal{D}_l=(\boldsymbol{x}^i,\boldsymbol{y}^i)_{i=1}^n$ 和 m 个无标记的训练样例 $\mathcal{D}_u=\{\boldsymbol{x}^j\}_{j=1}^m$，通常 m 远大于 n，这意味着无标记样本的数量远大于有标记样本的数量。本文的目标是基于标记数据集 $\mathcal{D}_l$ 和未标记数据集 $\mathcal{D}_u$ 训练一个模型 $f(\boldsymbol{x};\theta):\boldsymbol{x}\rightarrow\boldsymbol{y}$，其中 θ 是模型参数。为了符号的简单表示，在以下内容中省略符号 θ。给定一个未见过的实例 $\boldsymbol{u}\in\mathcal{X}$，该模型的概率分布预测为 $f(\boldsymbol{u})$，而 $f_k(\boldsymbol{u})$ 是第 k 个类的预测概率。

3.2 系统概述

本文方法框架如图 1 所示。首先本文方法通过编码器 $E:R^{d}\rightarrow R^{d_z}$ 将输入 $\boldsymbol{x}$ 编码为特征表示 $\boldsymbol{z}\in R^{d_z}$，其中 d_z 是特征表示 $\boldsymbol{z}$ 的维度。另外，$\boldsymbol{x}$ 根据数据集特征类型分为数值型和分类型。分类型为整数范围，固定为 0 或 1，不需要额外处理；数值型每个特征为浮点数，先对其进行最大最小标准化至 0 到 1 之间，再传给编码器 E。然后，将在 3.2 节中提出的隐式恒标数据增强模块应用于 $\boldsymbol{z}$，得到带噪声的特征表示 $\tilde{\boldsymbol{z}}$。同时，对所有标记样本 $\{\boldsymbol{x}\}^i$ 编码并选取当前样本特征表示 $\boldsymbol{z}$ 的 K 近邻 $\{\boldsymbol{z}\}^i$。3.3 节中提到的 K 近邻注意力机制模块使用特征表示 $\{\boldsymbol{z}\}^i$ 和对应的标记 $\{\boldsymbol{y}\}^i$ 来计算注意力分数 $\boldsymbol{s}$，该注意力分数可用于丰富当前输入特征表示 $\boldsymbol{z}$。最后，本文方法通过解码器 $D:R^{d_z}\rightarrow R^{d_z}$ 和分类器 $C:R^{d_z}\rightarrow R^{t}$ 进行预测。

对于模型训练，本文设计了监督损失 $\mathcal{L}_{\text{sup}}$ 和一致性损失 $\mathcal{L}_{\text{cons}}$。传统的监督多标记学习方法使用二元交叉熵（BCE）损失[19]，它将多标记分类任务拆分为多个单标记分类任务。然而，BCE 损失通常存在正负不平衡问题，为了缓解这个问题，采用了非对称损失（ASL）[20]。ASL 是焦点损失的一种变体，对正例和负例具有不同的焦点参数。因此，监督损失可以定义为：

$$\begin{aligned}\mathcal{L}_{\text{sup}}&=\mathcal{L}(f(\boldsymbol{x}),\boldsymbol{y})\\&=\sum_{k=1}^{t}y_k l_1(f_k(x))+(1-y_k)l_0(f_k(x))\end{aligned}\tag{1}$$

其中，$\mathcal{L}(f(\boldsymbol{x}),\boldsymbol{y})$ 是 ASL 损失；$l_1(f_k)=-(1-f_k)^{\lambda_1}\log(f_k)$ 和 $l_0(f_k)=-(f_k)^{\lambda_0}\log(1-f_k)$ 表示在正标记和负标记上计算的损失，λ_1 和 λ_0 是正负聚焦参数。对于未标记数据集 D_u 中大量无法获得其对应标记来直接计算 ASL 损失的样本，一种直观的策略是利用模型的预测对未标记样本进行伪标记，从而为这些样本分配伪标记来指导训练过程。于是，一个未标记样本的监督损失 $\mathcal{L}_{\text{sup}}$ 可以通过 3.4 节中提到的 $\mathcal{L}(f(\boldsymbol{x}),\hat{\boldsymbol{y}})$ 及其伪标记 $\hat{\boldsymbol{y}}$ 来计算。

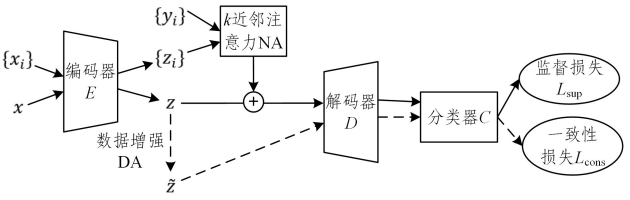


图 1 针对多标记表格数据的半监督学习方法框架

Fig. 1 Framework of semi-supervised learning for multi-label abular data

一致性损失 $\mathcal{L}_{\text{cons}}$ 旨在鼓励模型 f 在特征表示 $\boldsymbol{z}$ 被扰动为 $\tilde{\boldsymbol{z}}$ 时输出相似的标记概率分布。形式上，一致性损失如式（2）所示：

$$\mathcal{L}_{\text{cons}}=KL(C(D(\boldsymbol{z}+\boldsymbol{s}))\parallel C(D(\tilde{\boldsymbol{z}})+\boldsymbol{s}))\tag{2}$$

其中，$KL(\cdot\parallel\cdot)$ 是计算 KL 散度的函数，$\boldsymbol{s}$ 是在 3.3 节中计算的注意力分数。

基于监督损失和一致性损失，本文通过最小化目标函数 $\mathcal{L}_{\text{final}}$ 来训练预测模型 f：

$$\mathcal{L}_{\text{final}}=\mathcal{L}_{\text{sup}}+\beta\mathcal{L}_{\text{cons}}\tag{3}$$

其中，β 是一致性损失权重。然后，通过随机梯度下降（SGD）方法[21]优化模型参数。

3.3 隐式恒标数据增强

在半监督训练框架中，为了让训练出来的模型有更强的泛化能力，经常会使用数据增强的方法，比如在特征层面进行随机 mask 替换、在隐藏层进行两个样本的随机混合等，但直接操作特征和标记可能会破坏样本的原始信息，导致模型最终性能下降。

为了解决这个问题，本文提出在编码器提取特征表示后，按照式（4），在表示空间中随机添加高斯噪声，以实现隐式恒标数据增强。具体而言，$\tilde{\boldsymbol{z}}$ 的生成过程为：

$$\tilde{\boldsymbol{z}}=\boldsymbol{z}+\boldsymbol{m}\odot\boldsymbol{n}\tag{4}$$

其中，$\boldsymbol{m}=[m_1,\cdots,m_{d_z}]^{\mathrm{T}}\in\{0,1\}^{d_z}$ 是从伯努利分布 $\text{Bern}(\cdot\mid p_m)$ 中以概率 p_m 采样的随机掩码向量，该概率控制被掩盖和加噪的特征的比例；$\boldsymbol{n}\in R^{d_z}$ 是从高斯分布 $\mathcal{N}(0,\sigma_n)$ 中采样的随机噪声，σ_n 是控制噪声大小的超参数。随后将噪声添加到特征表示 $\boldsymbol{z}$，以获得加噪特征表示 $\tilde{\boldsymbol{z}}$。

基于上述隐式恒标数据增强方法，可以在不明显破坏原始特征和标记的情况下对样本进行扰动。

3.4 K 近邻注意力机制

在 ML-KNN 等多标记分类任务的先前工作中，经典的思路是利用从标记的邻近实例中获得的统计信息，该统计信息源自流行的 K 近邻（KNN）算法[22]。这表明使用标记样本的近邻信息对于改进多标记学习非常重要。

为了更好地利用近邻信息，本文提出了 K 近邻注意力机制模块。首先，直接使用经典算法 KNN 在标记数据集 $\mathcal{D}_l$ 中找到目标样本的 K 个最近邻居。但是，本文是在特征表示空间中搜索邻居而不是在初始特征空间中搜索邻居，这意味着不仅对目标样本进行了编码，还对整个标记数据集实例进行了编码。考虑到计算成本，标记数据集相对较小，可以在编码过程中跳过对它们的梯度计算。因此，可以得到 K 个近邻特征表示 $\{\boldsymbol{z}\}^i$ 及其对应的标记 $\{\boldsymbol{y}^i\}$。

然后，利用特征表示 $\boldsymbol{z}$、近邻特征表示 $\{\boldsymbol{z}^i\}$ 及其对应的标记 $\{\boldsymbol{y}^i\}$ 计算注意力得分 $\boldsymbol{s}$，如下所示：

$$\boldsymbol{s}=\text{softmax}\left(\frac{W_q(\boldsymbol{z})\cdot W_k(\{\boldsymbol{z}^i\}^{\mathrm{T}})}{\sqrt{d_l}}\right)\cdot W_v(\{\boldsymbol{y}^i\})\tag{5}$$

其中，$W_q:R^{d_z}\rightarrow R^{d_l}$，$W_k:R^{d_z}\rightarrow R^{d_l}$ 和 $W_v:R^{t}\rightarrow R^{d_z}$ 均为线性层。将特征表示 $\boldsymbol{z}$ 输入至 W_q，可得到维度 d_l 的查询；将近邻特征表示 $\{\boldsymbol{z}^i\}$ 输入至 W_k，可得到维度 d_l 的键；将对应的标记输入至 W_v，可得到维度 d_z 的值。因为一共有 K 个近邻，所以 $W_k(\{\boldsymbol{z}^i\})$ 结果为大小 $K\times d_l$ 的矩阵，$W_v(\{\boldsymbol{y}^i\})$ 结果为大小 $K\times d_z$ 的矩阵。故查询对应的矩阵与键对应的经过转置后的矩阵相乘的结果维度为 $1\times K$。softmax（·）表示一个 softmax 函数，它将查询和键相乘的结果映射到 0 到 1 之间，以获得值的权重，其维度不发生变化，仍是 $1\times K$。最后，和值对应的矩阵相乘得到结果注意力分数 $\boldsymbol{s}$，维度为 $1\times d_z$。将带有邻居信息的分数 $\boldsymbol{s}$ 添加到特征表示 $\boldsymbol{z}$，再送到解码器 D 进行后续预测。

3.5 伪标记

为了在半监督学习环境中充分利用未标记数据集$\mathcal{D}_u$，采用一种直观的方法，根据模型输出为未标记实例分配伪标记。本文保留一个教师模型来预测未标记样本，并使用正在训练的学生模型的指数滑动平均(EMA)更新教师模型参数，如下所示：

$$\theta_t=\gamma*\theta_{t-1}+(1-\gamma)*\theta \tag{6}$$

其中，θ_t是训练步骤t时的教师模型参数，θ是每轮训练步骤更新后的最新学生模型参数，γ是平滑系数超参数。

然后，将得到的未标记的训练样本在数据增强后输入给教师模型的输出，以获得它们的伪标记$\hat{\boldsymbol{y}}$。为了获得更可靠的伪标记，本文应用 CAP 方法。于是，可以计算后续的监督损失$\mathcal{L}_{\text{sup}}$。

4 实验

4.1 实验设置

为了全面评估本文方法，使用 10 个具有多样化属性的多标记表格数据集，这些数据集的属性如表 1 所列。对于数据集$\mathcal{D}$，其属性包括样本数量($|\mathcal{D}|$)、特征数量($dim(\mathcal{D})$)、类标记数量($L(\mathcal{D})$)、特征类型($F(\mathcal{D})$)、标记基数($LCard(\mathcal{D})$，即每个实例的平均标记数)。对于每个数据集，随机选择占总样本数比例为 $p=0.05$ 的具有完整标记的样本，而其余样本则没有任何监督信息。

表 1 实验数据集属性

Table 1 Characteristics of experimental data sets

数据集名称	$\|\mathcal{D}\|$	$dim(\mathcal{D})$	$L(\mathcal{D})$	$F(\mathcal{D})$	$LCard(\mathcal{D})$
corel5k	5000	499	374	分类	3.522
rcv1-s1	6000	944	101	数值	2.880
Corel16k-s1	13766	500	153	分类	2.859
delicious	16105	500	983	分类	19.020
iaprtc12	19627	1000	291	数值	5.719
espgame	20770	1000	268	数值	4.686
mirflickr	25000	1000	38	数值	4.716
tmc2007	28596	981	22	分类	2.158
mediamill	43907	120	101	数值	4.376
bookmarks	87856	2150	208	分类	2.028

按照先前的工作[23]，本文采用了 6 种广泛用于多标记分类的评估指标，包括平均精度(Average precision)、宏平均 AUC(Macro-averaging AUC)、汉明损失(Hamming Loss)、单次错误(One-error)、覆盖率(Coverage)和排序损失(Ranking Loss)。

将本文方法与针对表格数据设计的 6 种方法进行了比较. 每种方法的详细信息如下。

1)DNN：作为基本的监督学习基准，该模型只使用标记数据集$\mathcal{D}_l$进行训练。

2)VIME：一种基于一致性正则化的半监督学习方法，鼓励对原始样本和扰动样本的预测相似。扰动过程首先根据掩码分数p_m从伯努利分布生成掩码向量$\boldsymbol{m}$，用于破坏实例。然后，当m_j为 1 时，x_j被从$\mathcal{D}_l$和$\mathcal{D}_u$中的相同特征分布中采样替换。

3)SDAT[24]：一种基于一致性正则化的半监督学习方法，用于最小化未标记实例的输出与其增强输出之间的 KL 散度。增强过程在变分自动编码器的潜在空间中完成，通过从潜在编码$q(\boldsymbol{z}|\boldsymbol{x})$的后验分布中进行采样，然后通过解码器$p(\hat{\boldsymbol{x}}|\boldsymbol{z})$映射采样的$\hat{\boldsymbol{z}}$。

4)CUTMIX：一种应用于输入空间上的数据增强方法，可以应用于本文的设置。给定同一批次中的两个输入$\boldsymbol{x}_i,\boldsymbol{x}_j\in R^d$，随机选择$\lfloor\lambda\times k\rfloor$个索引并交换与所选索引相对应的$\boldsymbol{x}_i$，$\boldsymbol{x}_j$的值。标记混合过程为$\boldsymbol{y}_i=(1-\lambda)\boldsymbol{y}_i+\lambda\boldsymbol{y}_j$，其中$\lambda\sim\text{Beta}(\alpha,\alpha)$，$\alpha$是超参数。

5)SCARF：一种基于对比学习的方法，引入了应用于输入空间的数据增强技术。基于掩码比率α从伯努利分布生成的掩码向量$\boldsymbol{m}$用于确定掩码位置。当m_j为 1 时，x_j被替换为从该j个特征的经验边际分布中随机抽取的$\hat{x}_j$。

6)HiddenMix：一种应用于潜在空间的数据增强方法，可以应用于本文的设置。给定 Tokenizer 生成的同一批次中的两个$\boldsymbol{T}_i,\boldsymbol{T}_j\in R^{k\times d}$，增强的$\hat{\boldsymbol{T}}_i$通过公式$\hat{\boldsymbol{T}}_i=\boldsymbol{S}\odot\boldsymbol{T}_i+(\mathbb{1}-\boldsymbol{S})\odot\boldsymbol{T}_j$获得。系数矩阵$\boldsymbol{S}$和全一矩阵$\mathbb{1}$的大小为$k\times d$，$\boldsymbol{S}=[s_1,s_2,\cdots,s_k]^{\mathrm{T}}$，其中所有向量$\boldsymbol{s}_h\in R^d(h=1,2,\cdots,k)$都相同，并且有$\lfloor\lambda\times k\rfloor$个元素随机选择 1，其余元素为 0。标记混合过程为$\boldsymbol{y}_i=(1-\lambda)\boldsymbol{y}_i+\lambda\boldsymbol{y}_j$，其中$\lambda\sim\text{Beta}(\alpha,\alpha)$，$\alpha$是超参数。

采用与 DELA[15]相同的编码器和解码器架构来实现本文方法。具体来说，编码器E是一个具有 ReLU 激活的 4 层全连接神经网络，其中隐藏层维度设置为[256；512；256]；解码器D是一个由 256 个神经元组成的全连接线性层；分类器C是一个由 512 个神经元组成的全连接线性层。对于伪标记方法 CAP 中的超参数，考虑到数据集中正负类标记不平衡，存在大量负类标记，使用推荐设置阈值均为 1。对于网络优化，采用 AdamW 优化器[25]和单周期策略调度器[26]，最大学习率为 0.0001。所有数据集的预热训练轮数均设置为 20。

4.2 对比结果

表 2 和表 3 列出了所有指定评估指标的综合实验结果。从表中可以看出：

1)在 10 个数据集的所有评估指标中，本文方法在 80%的情况下取得了最佳性能，在 16.7%的情况下取得了第二好的性能。

2)与监督学习下仅使用标记数据的 DNN 方法相比，本文方法在 93.3%的情况下在各种评估标准上取得了改进。这证明对于半监督学习，本文方法可以充分利用未标记数据来增强模型性能。

3)与 CutMix 和 SCARF 在原始特征上应用数据增强不同，本文方法以及 SDAT 和 HiddenMix 等方法在表示空间中添加了噪声，并在所有数据集和评估指标中显示出更好的结果。这证明在潜在空间中执行数据增强非常重要，可以避免破坏原始特征之间的关系。

4)本文方法在所有评估指标上都比其他基于一致性正则化的方法(如 VIME 和 SDAT)取得了更优或相当的性能。这些结果一致证明了本文的近邻注意力设计对于利用标记数据的有效性。

上述结果有力地验证了本文的多标记表格数据半监督学习方法的有效性。

表 2 每种比较方法在平均精度、宏平均 AUC 和汉明损失方面的表现

Table 2 Performance of each comparing method in terms of average precision, Macro-averaging AUC and Hamming loss

数据集	平均精度↑						
	DNN	VIME	SDAT	CutMix	SCARF	HiddenMix	Ours
corel5k	0.1996	0.1995	0.2052	0.2020	0.2020	0.1990	**0.2190**
rcv1-s1	0.4518	0.4670	0.4617	0.2620	0.2619	0.4523	**0.5308**
Corel16k-s1	0.2379	0.2539	0.2613	0.2373	0.2368	0.2369	**0.2809**
delicious	0.2786	0.2944	0.2979	0.2346	0.2340	0.2862	**0.3163**
iaprtc12	0.2704	0.2808	0.2814	0.2206	0.2231	0.2775	**0.3042**
espgame	0.2320	0.2361	0.2359	0.1945	0.1952	0.2354	**0.2534**
mirflickr	0.5797	0.6010	0.5883	0.4882	0.4898	0.5887	**0.6099**
tmc2007	0.7099	**0.7698**	0.7332	0.5661	0.5662	0.7076	0.7632
mediamill	**0.7067**	0.6968	0.7047	0.6674	0.6673	0.6987	0.7049
bookmarks	0.3185	0.3530	0.3734	0.2536	0.2544	0.3267	**0.3879**
数据集	宏平均 AUC↑						
	DNN	VIME	SDAT	CutMix	SCARF	HiddenMix	Ours
corel5k	0.5071	0.5028	0.5110	0.5275	0.5280	0.5037	**0.5547**
rcv1-s1	0.7419	0.7615	0.7630	0.5736	0.5822	0.7382	**0.7851**
Corel16k0s1	0.5098	0.5744	0.5538	0.5086	0.5130	0.5053	**0.6120**
delicious	0.6833	0.6854	0.6828	0.5934	0.5959	0.6873	**0.7058**
iaprtc12	0.7139	0.7379	0.7350	0.6249	0.6262	0.7332	**0.7571**
espgame	0.6483	0.6543	0.6588	0.5560	0.5566	0.6455	**0.6761**
mirflickr	0.7584	0.7678	0.7263	0.6356	0.6329	0.7627	**0.7704**
tmc2007	0.8261	**0.8894**	0.8531	0.6003	0.6031	0.8240	0.8712
mediamill	**0.7416**	0.7297	0.7411	0.6937	0.6890	0.7206	0.7288
bookmarks	0.7656	0.8061	**0.8159**	0.6743	0.6714	0.7752	0.8093
数据集	汉明损失↓						
	DNN	VIME	SDAT	CutMix	SCARF	HiddenMix	Ours
corel5k	0.0481	0.0546	**0.0093**	0.0509	0.0505	0.0499	0.0094
rcv1-s1	0.0399	0.0773	0.0288	0.2185	0.2122	0.0549	**0.0284**
Corel16k0s1	0.1071	0.0954	**0.0189**	0.1102	0.1077	0.1092	0.0197
delicious	0.0895	0.0943	**0.0186**	0.1183	0.1122	0.0855	0.0187
iaprtc12	0.1142	0.1145	0.0195	0.1306	0.1285	0.1007	**0.0194**
espgame	0.0919	0.0937	0.0178	0.0993	0.0950	0.0855	**0.0177**
mirflickr	0.3700	0.4035	**0.1121**	0.5441	0.5455	0.3344	0.1125
tmc2007	0.1057	0.1102	0.0718	0.4895	0.4724	0.1077	**0.0689**
mediamill	0.1382	0.1514	0.0318	0.1824	0.1869	0.1434	**0.0316**
bookmarks	0.0528	0.0430	**0.0089**	0.0482	0.0498	0.0515	0.0094

注：↑(↓)表示值越大(越小)，性能越好；最佳结果以粗体突出显示，次优结果以下划线突出显示。

表 3 每种比较方法在单次错误、覆盖率和排序损失方面的表现

Table 3 Performance of each comparing method in terms of one-error, coverage and ranking loss

数据集	单次错误↓						
	DNN	VIME	SDAT	CutMix	SCARF	HiddenMix	Ours
corel5k	0.7780	0.7780	0.7720	0.7840	0.7720	0.7840	**0.7500**
rcv1-s1	0.5250	0.5258	0.5183	0.7558	0.7617	0.5450	**0.4675**
Corel16k0s1	0.8006	0.7607	0.7534	0.7905	0.7919	0.7984	**0.7168**
delicious	0.4502	0.4340	0.4331	0.5377	0.5299	0.4381	**0.4114**
iaprtc12	0.6014	0.5866	0.5789	0.6757	0.6689	0.5845	**0.5542**
espgame	0.6788	0.6822	0.6796	0.7414	0.7376	0.6784	**0.6673**
mirflickr	0.3856	0.3668	0.3800	0.5348	0.5388	0.3752	0.3464
tmc2007	0.3206	**0.2615**	0.3124	0.4385	0.4388	0.3206	0.2750
mediamill	**0.1718**	0.1866	0.1779	0.2069	0.2055	0.1768	0.1742
bookmarks	0.7091	0.6851	0.6592	0.7899	0.7871	0.7038	**0.6449**
数据集	覆盖率↓						
	DNN	VIME	SDAT	CutMix	SCARF	HiddenMix	Ours
corel5k	0.4132	0.4109	0.4136	0.4174	0.4143	0.4099	**0.3983**
rcv1-s1	0.2611	0.2392	0.2485	0.3771	0.3728	0.2602	**0.2018**
Corel16k0s1	0.3974	0.5744	0.3789	0.3945	0.3960	0.4009	**0.3587**
delicious	0.6320	0.6155	0.6292	0.6963	0.6981	0.6239	**0.6058**
iaprtc12	0.4245	0.3964	0.3944	0.4887	0.4887	0.4089	**0.3778**
espgame	0.4760	0.4661	0.4617	0.5290	0.5289	0.4736	**0.4327**
mirflickr	0.3532	0.3343	0.3494	0.4031	0.4012	0.3480	**0.3264**
tmc2007	0.1998	0.1557	0.1727	0.3016	0.3018	0.2052	**0.1518**
mediamill	**0.1677**	0.1736	0.1685	0.1948	0.1989	0.1729	0.1702
bookmarks	0.2638	0.2121	0.2051	0.3219	0.3234	0.2546	**0.1890**

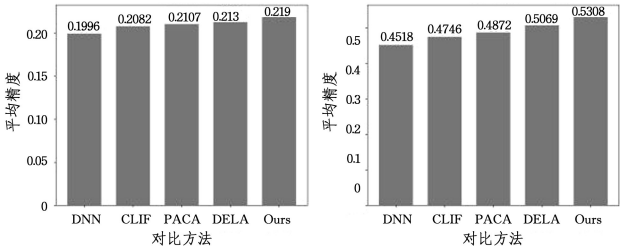
（续表）

数据集	排序损失↓						
	DNN	VIME	SDAT	CutMix	SCARF	HiddenMix	Ours
corel5k	0.1815	0.1817	0.1785	0.1821	0.1812	0.1806	**0.1729**
rcv1-s1	0.1337	0.1143	0.1196	0.1899	0.1886	0.1322	**0.0881**
Corel16k0s1	0.2125	0.2143	0.1999	0.2109	0.2116	0.2143	**0.1872**
delicious	0.1553	0.1428	0.1399	0.1669	0.1679	0.1521	**0.1328**
iaprtc12	0.1571	0.1472	0.1479	0.1885	0.1892	0.1549	**0.1368**
espgame	0.2054	0.1978	0.2037	0.2289	0.2293	0.2007	**0.1786**
mirflickr	0.1467	0.1328	0.1381	0.1769	0.1763	0.1428	**0.1293**
tmc2007	0.0997	0.0691	0.0819	0.1743	0.1743	0.1036	**0.0678**
mediamill	**0.0501**	0.0537	**0.0501**	0.0606	0.0619	0.0524	0.0503
bookmarks	0.1918	0.1476	0.1414	0.2345	0.2363	0.1822	**0.1279**

注：↑（↓）表示值越大（越小），性能越好；最佳结果以粗体突出显示，次优结果以下划线突出显示。

4.3 消融实验

在本节中，首先对比使用 3 种先进的多标记分类方法 CLIF[27]、PACA[28]、DELA 配合 EMA 模型和 CAP 伪标记方法的结果，如图 2 所示。可以看到，本文方法可以达到最佳性能并优于已有工作，说明了该方法的有效性。

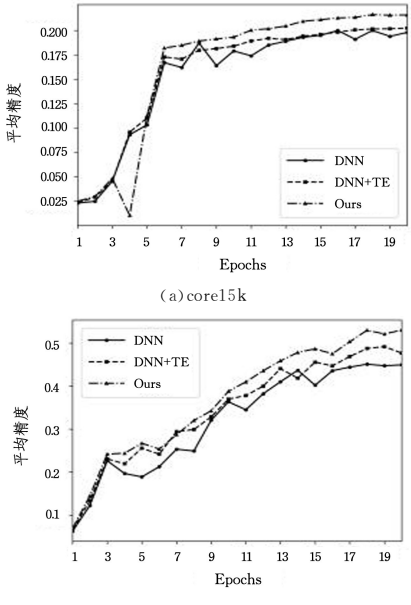


（a）corel5k （b）rcv1-s1

图 2 不同多标记方法配合伪标记方法对比

Fig. 2 Comparison of different multi-label methods with pseudo labeling

其次，对比了不使用伪标记、使用暂时集成（TemporalEnsembling，TE）和本文方法在训练过程中模型的测试平均精度收敛情况，结果如图 3 所示。可以看出，本文方法的收敛速度更快且性能更好。



（a）corel5k

（b）rcv1-s1

图 3 不同伪标记方法训练模型测试精度对比

Fig. 3 Comparison of test accuracy of training models using different pseudo-labeling methods

此外，还评估了本文方法的不同部分的效果，例如隐式恒标数据增强（DA）和 K 近邻注意力机制（NA）。删除数据增强部分和一致性损失作为 w/o DA，删除 K 近邻注意力机制部分作为 w/o NA。

在两个代表性数据集上的平均精度结果如表 4 所列，结果证明了本文所设计的数据增强和近邻注意力部分都对最终模型性能有益，并且数据增强带来了更多提升。

表 4 数据增强（DA）和邻域注意（NA）等不同部分在平均精度方面的表现

Table 4 Performance of different parts such as data augmentation (DA) and neighbor attention(NA) in terms of average precision

数据集	平均精度↑		
	Ours	w/o DA	w/o NA
corel5k	**0.2190**	0.2133	0.2170
rcv1-s1	**0.5307**	0.4605	0.4885

为探究超参数 β 对于平均精度的影响，在 3 个代表性数据集上进行了实验，结果如图 4 所示。其中 $\beta=0$ 时性能均为最差，证明了一致性损失的有效性。

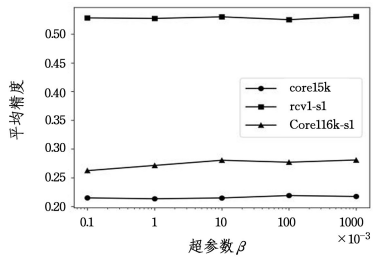


图 4 超参数 β 对平均精度的影响

Fig. 4 Effect of hyper-parameter β on average precision

最后，讨论了改变 K 近邻注意力机制（NA）中的超参数 K 所带来的影响，在两个代表性数据集上进行了实验，结果如图 5 所示，可以看到，K 值的改变并不会引起性能的剧烈变化，说明了该参数的鲁棒性。

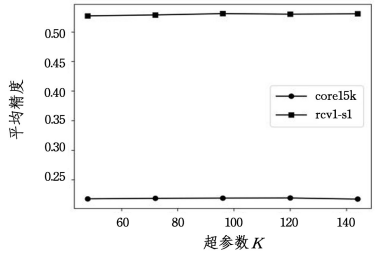


图 5 超参数 K 对平均精度的影响

Fig. 5 Effect of hyper-parameter K on average precision

结束语 本文提出针对多标记表格数据的半监督学习，考虑到传统的数据增强方法并不适用于此，设计了一种新的多标记表格数据隐式恒标数据增强方法，避免破坏原始数据结构。此外，通过 K 近邻注意力机制充分利用近邻信息，并通过 EMA 模型和 CAP 方法提高伪标记质量。值得补充的是，在半监督场景中，本文方法只需要少量已标注数据就可以有效解决医疗金融领域中，结合联邦学习因隐私问题导致的数据量不足。在 10 个公共数据集上进行的大量实验结果证明了本文方法的有效性。未来，计划在半监督场景中为表格数据设计更具有针对性的伪标记框架。

参 考 文 献

[1] SOMVANSHI S, DAS S, JAVED S A, et al. A survey on deep tabular learning[J]. arXiv: 2410.12034, 2024.

[2] TAREKEGN A N, ULLAH M, CHEIKH F A. Deep learning for multi-label learning: a comprehensive survey[J]. arXiv: 2401.16549, 2024.

[3] OUALI Y, HUDELOT C, TAMI M. An overview of deep semi-supervised learning[J]. arXiv: 2006.05278, 2020.

[4] LEE D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]// Workshop on Challenges in Representation Learning. New York: ICML, 2013: 896.

[5] XIE Q, DAI Z, HOVY E, et al. Unsupervised data augmentation for consistency training[J]. Advances in Neural Information Processing Systems, 2020, 33: 6256-6268.

[6] LAINE S, AILA T. Temporal ensembling for semi-supervised learning[J]. arXiv: 1610.02242, 2016.

[7] JIA S, WANG P, JIA P, et al. Research on data augmentation for image classification based on convolution neural networks [C]// 2017 Chinese Automation Congress (CAC). Piscataway, NJ: IEEE, 2017: 4165-4170.

[8] SHORTEN C, KHOSHGOFTAAR T M, FURHT B. Text data augmentation for deep learning[J]. Journal of big Data, 2021, 8(1): 101.

[9] LAINE S, AILA T. Temporal ensembling for semi-supervised learning[J]. arXiv: 1610.02242, 2016.

[10] YOON J, ZHANG Y, JORDON J, et al. Vime: Extending the success of self-and semi-supervised learning to tabular domain [J]. Advances in Neural Information Processing Systems, 2020, 33: 11033-11043.

[11] BAHRI D, JIANG H, TAY Y, et al. Scarf: Self-supervised contrastive learning using random feature corruption[J]. arXiv: 2106.15147, 2021.

[12] SOMEPALLI G, GOLDBLUM M, SCHWARZSCHILD A, et al. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training[J]. arXiv: 2106.01342, 2021.

[13] CHEN J, YAN J, CHEN Q, et al. Excelformer: A neural network surpassing gbdts on tabular data[J]. arXiv: 2301.02819, 2023.

[14] ZHANG M L, ZHOU Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[15] HANG J Y, ZHANG M L. Dual perspective of label-specific feature learning for multi-label classification[J]. ACM Transactions on Knowledge Discovery from Data, 2024, 19(1): 1-30.

[16] LI G Z, YANG J Y, LU W C, et al. Improving prediction accuracy of drug activities by utilising unlabelled instances with feature selection[J]. International Journal of Computational Biology and Drug Design, 2008, 1(1): 1-13.

[17] XIE M K, XIAO J, LIU H Z, et al. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning[J]. Advances in Neural Information Processing Systems, 2023, 36: 25731-25747.

[18] LIU B, XU N, FANG X, et al. Correlation-induced label prior for semi-supervised multi-label learning[C]// Forty-first International Conference on Machine Learning. 2024.

[19] GOODFELLOW I, BENGIO Y, COURVILLE A, et al. Deep learning[M]. Cambridge: MIT press, 2016.

[20] RIDNIK T, BEN-BARUCH E, ZAMIR N, et al. Asymmetric loss for multi-label classification[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 82-91.

[21] AMARI S. Backpropagation and stochastic gradient descent method[J]. Neurocomputing, 1993, 5(4/5): 185-196.

[22] PETERSON L E. K-nearest neighbor [J]. Scholarpedia, 2009, 4(2): 1883.

[23] ZHANG M L, ZHOU Z H. A review on multi-label learning algorithms[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(8): 1819-1837.

[24] FANG J, TANG C, CUI Q, et al. Semi-supervised learning with data augmentation for tabular data[C]// Proceedings of the 31st ACM International Conference on Information & Knowledge Management. New York: ACM, 2022: 3928-3932.

[25] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[J]. arXiv: 1711.05101, 2017.

[26] DEVRIES T, TAYLOR G W. Improved regularization of convolutional neural networks with cutout[J]. arXiv: 1708.04552, 2017.

[27] HANG J Y, ZHANG M L. Collaborative learning of label semantics and deep label-specific features for multi-label classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(12): 9860-9871.

[28] HANG J Y, ZHANG M L, FENG Y, et al. End-to-end probabilistic label-specific feature learning for multi-label classification [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 6847-6855.

GE Zeqing, born in 2000, postgraduate. His main research interests include multi-label learing and semi-supervised learning.

HUANG Shengjun, born in 1987, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 42916S). His main research interests include machine learning and pattern recognition.

（责任编辑：何杨）