计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330238
60(8): 1740−1753, 2023

面向特征演变环境的标记噪声鲁棒学习算法

张震宇　姜　远
（南京大学计算机科学与技术系　南京　210023）
（zhangzy@lamda.nju.edu.cn）

Label Noise Robust Learning Algorithm in Environments Evolving Features

Zhang Zhenyu and Jiang Yuan
（*Department of Computer Science and Technology*, *Nanjing University*, *Nanjing* 210023）

Abstract　In real-world applications, data are often collected in the form of a stream, with features that can evolve over time. For instance, in the environmental monitoring task, features can be dynamically vanished or augmented due to the existence of expired old sensors and deployed new sensors. Additionally, besides the evolvable feature space, the labels potentially contain noise. When feature space evolves and data conceal inaccurate labels at the same time, it is quite challenging to design algorithms with guarantees, particularly theoretical understandings of generalization ability. To address this difficulty, we propose a new discrepancy measure for noisy labeled data with evolving feature space, named the label noise robust evolving discrepancy. Using this measure, we present the generalization error analysis, and the theory motivates the design of a learning algorithm which is further implemented by deep neural networks. Empirical studies on synthetic data confirm the rationale of our discrepancy measure and extensive experiments on real-world tasks validate the effectiveness of our algorithm.

Key words　label noise; evolving feature space; weakly supervised learning; open-environment; robust learning

摘　要　在现实应用中，数据通常以流的形式不断积聚，数据的特征可能随时间而演变. 例如，在环境监测任务中，由于旧传感器达到使用寿命和新传感器的部署，数据特征可能会动态地消失或增加. 此外，除了可演变的特征空间，数据标记可能存在噪声. 当特征空间演变和数据标记带噪同时发生时，设计具有理论保障的学习算法，尤其是具备对算法泛化能力的理解是非常具有挑战性的. 为了应对这一挑战，提出了一种在特征演变环境中针对标记带噪数据的差异度量方法，称为容忍标记噪声的演变差异. 该差异度量启发了泛化误差分析，并根据泛化误差的理论分析设计了一种基于深度神经网络实现的学习算法. 合成数据上的实证研究验证了所提差异度量的合理性，而在现实应用任务上的实验则验证了所提算法的有效性.

关键词　标记噪声；特征演变环境；弱监督学习；开放环境；鲁棒学习

中图法分类号　TP181

传统的机器学习任务通常假设学习环境稳定不变，并且收集的数据标记是高质量的. 然而，在实际的应用场景中，数据在开放变化的环境中不断累积，而且数据标记往往是带有噪声的. 具体而言，随着学习环境的变化，数据的特征空间可能发生演变. 与此同时，数据的标记可能不准确. 这类问题在现实应用中广泛存在，并且具有重要意义. 以环境监测任务为例，研究者在野外部署传感器以监测种群信息. 传感器不断收集数据，每个传感器收集的数值对应着一个特征，所有传感器收集的数据组合在一起形成了

收稿日期：2023-03-31；**修回日期**：2023-06-12
基金项目：国家自然科学基金项目（62176117）
This work was supported by the National Natural Science Foundation of China (62176117).
通信作者：姜远（jiangyuan@nju.edu.cn）

数据的特征空间. 由于传感器使用寿命有限，因此研究人员需要用新的传感器替换损坏的传感器. 随着时间的推移，新的特征（新传感器）出现，旧的特征（到达使用寿命的传感器）消失，学习要素发生了变化，即数据的特征空间发生了演变. 与此同时，由于对数据样本的人工标注可能存在误差，数据标记可能存在噪声. 因此使得学习系统具备应对开放环境下特征空间演变和数据标记带噪的能力变得至关重要[1-2].

随着数据特征空间的不断演变，学习模型需要具备处理新特征空间数据的能力，以建立具有泛化能力的学习模型. 为此，建立新旧特征空间之间的关系，有效利用旧特征空间的历史数据是至关重要的. 以往的研究通常考虑特征演变环境中的数据流存在特征演变阶段[3-4]. 研究者们通过该阶段的数据建立新旧特征空间之间的联系，从而利用历史数据帮助学习新特征空间下的分类器. 以环境监测为例，如图 1 所示，在传感器到达使用寿命之前，研究者可以提前部署一批新的传感器用以替代即将失效的传感器. 在此阶段，新旧传感器同时收集数据，因此此时的数据具有新旧特征空间的表示，形成了数据演变阶段. 为了构建新特征空间上的分类器，研究者们提出了一种基于映射函数的解决思路[3]. 该方法利用演变阶段的数据学习映射函数，在新特征空间上学习分类器的同时，将数据映射回旧特征空间，并复用旧特征空间的模型，辅助构建新特征空间上的分类器.

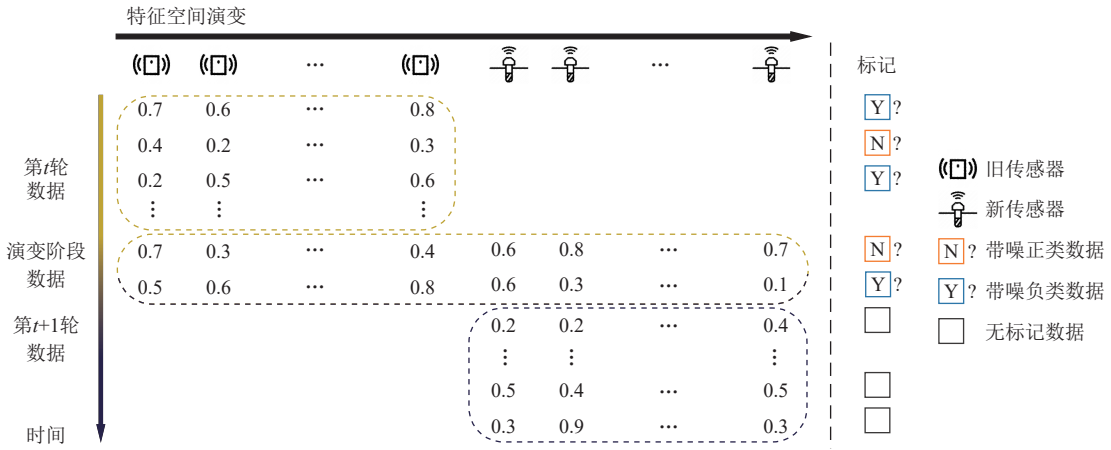


Fig. 1 Learning with noisy labeled data in environments with evolving features

图 1 特征演变环境下的标记带噪数据学习

除了数据特征空间的演变外，由于标注过程中可能存在人为误差，数据的标记往往受到噪声的干扰[5-6]. 以野生动物种群监测为例，某些动物可能难以进行准确区分，如猎豹和美洲虎、狼和土狼等[7]. 当数据特征空间发生演变并且数据标记存在噪声时，仅仅叠加处理特征演变的算法和处理标记噪声的算法无法很好地解决这一问题，尤其是很难获得在理论上具有良好泛化性能保障的分类器. 具体而言，为了应对这一难题，一种可行的方案是利用演变阶段的数据，首先学习新旧特征空间数据表示之间的映射方式，建立它们之间的联系. 其次，模型可以通过旧特征空间的带噪标记数据进行鲁棒学习，得到分类器. 最后，基于学习得到的映射函数，算法将新特征空间的数据映射到旧特征空间上进行预测. 然而，由于演变阶段的数据较少，通过演变阶段数据学习得到的映射函数通常无法完美地还原新特征空间数据在旧特征空间上的表示，因此这种方法的效果会受到映射函数学习质量的制约. 在对经过映射函数恢复出的特征表示进行预测时，算法会放大构建映射函数时所引入的误差，导致分类器性能下降. 另外，通过映射函数进行学习的启发式算法缺乏理论理解，难以获得具有泛化性能保障的分类器. 对于特征空间演变环境下的标记带噪数据流学习而言，目前还缺乏能够提供泛化性能保障的算法设计. 因此，设计一种有效的算法，能够同时处理特征空间的演变和数据标记上的噪声，并在新特征空间上获得具备泛化性能保障的分类器，成为了当前重要的研究课题.

本文首先形式化了一种特征演变环境下的标记带噪数据学习问题. 具体来说，本文考虑训练数据以流的形式不断累积的学习场景，数据特征空间发生演变，同时环境返回的数据标记存在噪声. 在这种情况下，学习者需要不断对新到达的无标记数据进行预测，并期望在新特征空间上获得具有泛化能力的分类器. 该问题在现实应用中经常遇到，但却没有得

到深入的研究. 针对这类问题，设计具有泛化性能保障的分类器十分重要且具有挑战性. 为了应对这个问题，本文旨在提出一种基于数据差异最小化技术[8-10]的算法来刻画分类器在新特征空间上的泛化能力. 然而，现有的数据差异最小化技术通常只考虑特征空间不变的数据，并且认为数据的标记是准确的，这在本文考虑的学习问题中并不适用. 因此，为了学习具有泛化能力的分类器，学习者需要刻画特征空间演变环境下不同特征空间数据的差异度量，并设计相应的具有理论保障的学习算法. 本文在数据标记存在噪声的情况下，利用演变阶段数据，建立了新旧特征空间之间的数据差异度量. 基于所提出的容忍标记噪声的数据差异度量，本文进一步导出了相应的泛化误差分析结果，并提出了一种基于深度神经网络的学习算法 LREDM(label-noise robust evolving discrepancy minimization).

本文的主要贡献包括 3 个方面：

1) 引入并形式化了特征演变环境下标记带噪数据的学习问题，尽管这类问题在现实应用中广泛存在，但却缺乏深入的研究；

2) 定义了容忍标记噪声的数据差异度量，并基于此进行了模型泛化性能的分析，给出了模型泛化误差界的理论保障；

3) 基于提出的理论设计了相应的基于深度神经网络实现的学习算法 LREDM，并且在现实应用的数据集上验证了所提算法的有效性.

1 相关工作

本文旨在研究特征演变环境下的弱监督学习. 在实际应用场景中，训练数据通常来自于开放且带噪的环境，因此，学习系统具备适应学习要素变化和标记带噪数据的能力是至关重要的[1-2]，这也是稳健机器学习[11]和学件[12]的关键要求. 本文主要关注 2 个方面：一是特征演变数据学习，二是标记带噪数据学习. 接下来，本节将从这 2 个方面概述本文的相关工作.

1.1 特征演变数据学习

早期的工作主要研究增量属性学习[12]和梯形数据学习[13]. 这些研究考虑数据样本及其特征维度不断增加的情况，学习者需要不断更新模型以应对新增的数据特征. 文献 [14] 提出了 OLSF(online learning with streaming features)算法，以解决数量与特征维度随时间增加的梯形数据问题. 当新特征维度到达时，OLSF 算法对不同特征空间执行不同的在线更新规则，使分类器不断适应新特征维度. 与这些研究不同的是，本文的研究问题考虑在新特征出现的同时旧特征可能消失，这为特征空间演变环境下的机器学习带来了新的挑战. 为了处理特征空间演变的数据流，一系列前沿的研究工作考虑建立新旧特征空间的联系，进而复用历史信息辅助对新特征空间数据进行预测和学习. 文献 [3] 的开创性工作考虑了数据演变阶段，其中新旧特征空间的数据同时存在. 文献 [3] 提出了 FESL(feature evolvable streaming learning)算法，通过演变阶段学习映射函数，恢复新特征空间上的数据在旧特征空间上的表示.FESL 算法在新旧特征空间上构建了 2 个学习模型，通过在线集成的方式进行预测和学习. 后续的工作对演变阶段映射方法的学习进行了拓展和探索[15]. 文献 [4] 进一步研究了这一问题，建立了新旧特征空间之间的演变差异，提出了 EDM(evolving discrepancy minimization)算法，最小化了分类器在新特征空间数据上的泛化误差. 文献 [16] 则考虑了特征继承性增减环境下的学习问题. 其中，新特征空间与旧特征空间的数据存在若干重叠特征，即部分特征表示在旧特征空间和新特征空间内同时存在. 文献 [16] 提出的 OPID(one-pass incremental and decremental learning)算法，通过将新特征空间数据映射回旧特征空间，使得旧特征空间数据得以复用进行学习.OFID(online classification algorithm with feature inheritably increasing and decreasing)算法在此基础上对特征继承性增减环境进行了深入研究[17]. 此外，一些研究者考虑了数据流中特征任意增减的情况. 例如，文献 [18] 提出的 OCDS(online learning from capricious data streams)算法和文献 [19] 提出的 OLVF(online learning from varying features)算法，将数据映射到共享的特征空间中进行统一学习. 除此以外，特征演变环境下的在线度量学习[20]、演变数据特征缺失学习[21]同样得到了研究者关注. 然而，大多数特征演变数据学习算法假设数据的标记是准确的，因此难以适应弱监督学习场景，尤其是当数据标记存在噪声时.

1.2 标记带噪数据学习

标记带噪数据学习旨在利用标记带噪数据训练出潜在无噪声数据分布上具有良好泛化能力的分类器. 自文献 [22−23] 的开创性研究，机器学习领域对标记带噪数据学习进行了深入广泛地研究. 本文着重研究类别相关的标记噪声，即噪声标记的形成机制仅与其真实所属类别有关.

文献 [6] 的研究表明，当噪声率已知时，通过设置适当的凸代理损失函数，可以在标记噪声类别下进行经验风险最小化，从而得到与无噪声数据分布上最优分类器一致的分类器. 文献 [24] 通过代理损失分解的方法，提出了样本标记中心平滑的方法，降低了标记噪声的负面影响. 然而，在实际应用中，学习问题往往面对噪声率未知的场景. 此时，估计噪声率或噪声转移矩阵是处理类别相关的均匀标记噪声的核心挑战之一. 研究者们提出了一系列假设来估计噪声转移矩阵，例如锚词条件[25]、锚点[26-27]和不可约性[28-29]. 具体而言，文献 [26] 的研究考虑了噪声率未知的情况，通过估计噪声率上界的形式，给出了设计凸代理损失函数的方案，并在实验中验证了其有效性. 文献 [27] 假设存在一些经过验证的准确数据，通过准确标记与无标记数据提取信息来估计噪声率. 文献 [28] 对标记带噪数据做出不可约假设，即存在一些锚点. 因此，学习者通过锚点从有噪声的标记数据中区分并恢复无噪声分布. 通过对数据分布做出一定假设，学习者从数据中估计噪声率或噪声转移矩阵，进而恢复潜在无噪声数据分布，从而获得具有泛化性能的学习器. 然而，大多数标记带噪数据学习方法集中在静态稳定环境中，一旦学习要素发生变化，例如特征空间的改变，导致这些方法往往无法直接适用.

2 预备知识

本节首先介绍特征演变环境下标记带噪数据学习的场景以及本文用到的符号和相关定义，然后对数据分布差异度量的基本概念及其应用进行回顾.

2.1 问题设定和相关符号

本节考虑特征空间演变环境中的标记带噪数据学习任务. 首先介绍该场景下的数据特点，接着给出相关符号和定义，最后给出问题的形式化.

与文献 [3] 的先驱性工作相似，本文假设数据在特征空间演变的过程中存在数据演变阶段. 如图 1 所示，学习者在旧传感器即将到达使用寿命之前可以提前部署一批新的传感器，防止发生旧传感器失灵后无法收集数据的问题. 因此，在提前布置好新的传感器之后，学习者可以收到少量具有新旧 2 个特征空间表示的数据连接先后 2 批相邻但不同特征空间的数据块. 与此同时，由于数据在收集过程中往往存在着噪声，数据的标记可能是错误的.

本文将上述场景形式化为在特征空间演变场景中的标记带噪数据的学习问题. 令 $\mathcal{X}_t \subseteq \mathbb{R}^{d_t}$ 表示第 t 轮数据的特征空间，学习者每轮首先收到样本集合 $\{x_i^t\}_{i=1}^{n_t}$，其中 $x_i^t \in \mathcal{X}_t$. 本文考虑特征空间发生演变，因此有 $d_t \neq d_{t+1}, t = 1, 2, \cdots, T$. 依照传统在线学习的设置，算法首先需要对样本进行预测，随后环境返回样本的带噪标记. 考虑本文的问题场景，研究二分类学习问题，即样本的类别空间 $\mathcal{Y} = \{-1, +1\}$，但是观测到的样本标记存在标记噪声. 对每个样本 x_i，本文将样本未被观测到的真实标记为 y_i，被观测到的噪声标记为 $\tilde{y}_i$. 对于每个真实标记为 y_i 的样本，它以一定概率以噪声标记 $\tilde{y}_i$ 被观测到，这一概率根据噪声率 $\rho_{\pm 1}$ 定义[6]，具体而言

$$Pr\left[\tilde{y} = +1 | y = -1\right] = \rho_{-1},$$

$$Pr\left[\tilde{y} = -1 | y = +1\right] = \rho_{+1}.$$

值得注意的是，噪声率 $\rho_{\pm 1}$ 对于算法而言是未知的.

在对 $t+1$ 轮的数据进行预测时，算法具有第 t 轮的标记带噪数据和第 $t+1$ 轮的待预测无标记数据. 因此，根据本文研究场景中的数据特点，每一时刻学习者拥有的数据可以被划分成 3 个部分：噪声标记数据、演变阶段数据、待预测数据：

1）来自第 t 轮特征空间的噪声标记数据集合，记作 $\tilde{L}_t^{\rho} = \{(x_1^t, \tilde{y}_1^t), \cdots, (x_{n_t}^t, \tilde{y}_{n_t}^t)\}, (x_i^t, \tilde{y}_i^t) \in \mathcal{X}_t \times \mathcal{Y}$.

2）演变阶段数据具有新旧 2 个特征空间 $\mathcal{X}_t$ 和 $\mathcal{X}_{t+1}$ 上的表示. 令 $U_t' = \{x_1^t, x_2^t, \cdots, x_k^t\}$ 表示基于旧特征空间表示的演变阶段数据，$U_{t+1}' = \{x_1^{t+1}, x_2^{t+1}, \cdots, x_k^{t+1}\}$ 表示基于新特征空间表示的演变阶段数据，其中 $x_i^t \in \mathcal{X}_t$，$x_j^{t+1} \in \mathcal{X}_{t+1}$.

3）来自第 $t+1$ 轮的新特征空间的待预测数据，记作 $U_{t+1} = \{x_1^{t+1}, x_2^{t+1}, \cdots, x_{n_{t+1}}^{t+1}\}$，其中 $x_j^{t+1} \in \mathcal{X}_{t+1}$.

值得注意的是，演变阶段并不会延续太久，因此有 $k \ll n_t$ 和 $k \ll n_{t+1}$. 以环境监测任务为例，提前部署新传感器以替换旧传感器的过程形成了数据演变阶段，这一阶段的数据数量通常少于旧特征空间的数据. 由于演变阶段样本较少，如果直接利用演变阶段数据的新特征空间上的表示进行学习，分类器会产生严重的过拟合现象. 为了缓解 $t+1$ 轮新特征空间上样本数量较少的问题，本文考虑利用第 t 轮旧特征空间的标记带噪样本辅助进行学习. 表 1 总结了本文中使用到的主要符号及其定义.

对于特征空间演变的标记带噪数据的学习问题，本文的目标是构建一个在新特征空间上具有泛化能力的分类器. 令分类器属于给定函数族 $\mathcal{G}$，其中每个函数 $g_t : \mathcal{X}_t \mapsto \mathbb{R}$. 考虑损失函数 $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ 和函数 $g \in \mathcal{G}$，本文将 $R_{\mathcal{D}}(g)$ 记作分类器的期望风险，

$$R_{\mathcal{D}}(g) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(g(x), y)\right] . \quad (1)$$

Table 1 Notations and Their Meaning

表 1 符号和对应含义

符号	含义
x_i	无标记样本
y_i	样本x_i的准确标记
$\tilde{y}_i$	样本x_i的带噪标记
$\tilde{L}_t^{\rho}$	第t轮标记带噪样本集合
U_t	第t轮待预测样本集合
$\mathcal{X}_t$	第t轮数据的特征空间
g	分类器
$\mathcal{G}$	分类器的集合
ρ_{-1}	负类样本噪声率
ρ_{+1}	正类样本噪声率
n_t	第t轮样本数
k	第t轮演变阶段样本数

$\hat{R}_{\mathcal{D}}(g)$记作分类器的经验风险,

$$\hat{R}_{\mathcal{D}}(g)=\frac{1}{n}\sum_{i=1}^{n}\ell(g(x_i),y_i)\ . \tag{2}$$

因此,本文的学习目标是对于每一轮收集得到的样本,构建在新特征空间上具有泛化性能的分类器,即在新特征空间上获得一个期望风险较小的分类器.

2.2 数据分布差异度量

为了应对特征空间的演变,本文将借助差异最小化技术设计算法.本节回顾这一领域的相关研究内容.对于2个来自不同分布的数据集,研究者们提出了多种差异度量方式,用以刻画它们之间的差异.基于所提的差异度量,研究者们设计对应算法,将数据差异最小化,进而可以在源数据(source data)上学习分类器,并部署到另一个目标分布(target distribution)对应的数据上.例如,利用KL散度刻画分布差异的学习方法KLIEP(Kullback-Leibler importance estimation procedure)算法[30],利用最大均值差异度量分布差异的学习方法KMM(kernel mean matching)算法[31]等.文献[8]基于所提的$\mathcal{H}\Delta\mathcal{H}$散度给出了一种广义的差异最小化算法的分析.后续由文献[9]拓展到了任意损失函数上.文献[32]进一步定义了$\mathcal{Y}$散度,并基于此给出了更紧的泛化误差界.这些基于差异的泛化误差界启发了差异最小化算法[10].

对于训练数据集和测试数据集分布不同的问题,学习者希望利用有标记的训练数据集训练在无标记的测试数据集上具有良好泛化能力的模型.假设训练数据与测试数据属于同一个特征空间,对于一个固定的分类器$g\in\mathcal{G}$而言,它在2个不同分布P和Q上进行迁移部署的质量可以通过分类器期望损失的差异$\left|R_P(g)-R_Q(g)\right|$来度量.一个直观的差异度量方式就是可以通过对未知的分类器和目标函数的差异取最大值,定义为:

$$disc(P,Q)=\max_{g,g'\in\mathcal{G}}\left|R_P(g;g')-R_Q(g;g')\right|,$$

其中$R_P(\cdot;g')$表示分类器在数据分布P上目标函数为g'的期望损失.

更进一步地,研究者们常会用$\mathcal{Y}$散度[32]来衡量数据分布之间的差异最大值,其定义为:

$$disc_{\mathcal{Y}}(P,Q)=\max_{g\in\mathcal{G}}\left|R_P(g;f_P)-R_Q\left(g;f_Q\right)\right|.$$

这一定义依赖于已知的目标函数f_P和f_Q,它们是分布差异$disc(P,Q)$的一种推广.

值得注意的是,这些方法通常假设数据分布在同一个特征空间内,因此不适用于本文考虑的特征空间发生演变的学习场景.

3 理论与方法

本节为特征空间演变环境下的标记带噪数据学习问题建立理论与方法.为了获得具有泛化性能保障的分类器,首先重写了分类器在新特征空间上的泛化误差,而其中的关键要素是在标记存在噪声的情况下刻画2个不同特征空间数据之间的差异.基于对分类器泛化能力的分析,本节进行了对应的算法设计,并通过深度神经网络进行了具体实现.

3.1 容忍标记噪声的演变差异

对于特征空间演变环境中的学习问题,由于演变阶段数据量较少,仅依赖演变阶段的标记数据来训练模型会容易出现严重的过拟合现象.因此,在这种情况下,利用旧特征空间上的标记数据变得至关重要.然而,由于新旧特征空间的维度不同,固定特征空间上的数据分布差异度量不再适用.此外,由于数据标记存在噪声,算法无法直接利用,因此学习者亟需建立容忍标记噪声的异质特征空间数据差异度量.

为了应对标记噪声,本文采用样本加权的方法恢复分类器在潜在无噪声数据上的真实损失.具体而言,针对每一轮的标记带噪数据集合$\tilde{L}_t^{\rho}$,本文定义其加权经验损失为

$$\hat{R}_t^{\rho}(g)=\frac{1}{n_t}\sum_{i=1}^{n_t}\alpha_i\cdot\ell(g(x_i),\tilde{y}_i), \tag{3}$$

其中样本权重α_i定义为

$$\alpha_i=\frac{Pr_{\mathcal{D}^{\rho}}(\tilde{y}_i|x_i)-\rho_{-\tilde{y}_i}}{\left(1-\rho_{\tilde{y}_i}-\rho_{-\tilde{y}_i}\right)Pr_{\mathcal{D}^{\rho}}(\tilde{y}_i|x_i)}. \tag{4}$$

基于式(1)的权重定义，考虑分类器在标记带噪数据上的加权损失，有引理 1 成立[6,26].

引理 1. 对于假设空间中的任一函数$g\in\mathcal{G}$，给定权重α，分类器在标记带噪数据上的加权期望损失等于该分类器在潜在无噪声数据上的真实损失，即

$$\mathbb{E}_{(x,\tilde{y})\sim\mathcal{D}^{\rho}}\left[\hat{R}_t^{\mathcal{D}^{\rho}}(g)\right]=R_t(g). \tag{5}$$

证明：根据对期望风险的概率拆分，有：

$$R_t(g)=\mathbb{E}_{(x,y)\sim D_t}\left[\ell(g(x),y)\right], \tag{6}$$

$$R_t(g)=\mathbb{E}_{(x,\tilde{y})\sim\mathcal{D}_t^{\rho}}\left[\frac{Pr[x,y]}{Pr[x,\tilde{y}]}\ell(g(x),y)\right], \tag{7}$$

$$R_t(g)=\mathbb{E}_{(x,\tilde{y})\sim\mathcal{D}_t^{\rho}}\left[\frac{Pr[y|x]}{Pr[\tilde{y}|x]}\ell(g(x),y)\right], \tag{8}$$

$$R_t(g)=\mathbb{E}_{(x,\tilde{y})\sim\mathcal{D}_t^{\rho}}\left[\frac{\Pr[\tilde{y}|x]-\rho_{-\tilde{y}}}{(1-\rho_{-1}-\rho_{+1})\Pr[\tilde{y}|x]}\ell(g(x),y)\right]. \tag{9}$$

其中式(6)是根据定义展开的，式(7)是改变数据分布后的概率重加权，式(8)的成立是因为标记噪声不改变样本特征的边际分布，式(9)的成立是根据噪声率的定义. 对于式(9)，考虑噪声正类标记的后验概率，有：

$$\begin{aligned}
&Pr[\tilde{y}=+1|x]=\\
&Pr[\tilde{y}=+1,y=+1|x]+Pr[\tilde{y}=+1,y=-1|x]=\\
&Pr[\tilde{y}=+1|y=+1,x]\,Pr[y=+1|x]+\\
&Pr[\tilde{y}=+1|y=-1,x]\,Pr[y=-1|x]=\\
&Pr[\tilde{y}=+1|y=+1]\,Pr[y=+1|x]+\\
&Pr[\tilde{y}=+1|y=-1]\,Pr[y=-1|x]=\\
&(1-\rho_{+1})Pr[y=+1|x]+\rho_{-1}(1-Pr[y=+1|x])=\\
&(1-\rho_{+1}-\rho_{-1})Pr[y=+1|x]+\rho_{-1}.
\end{aligned}$$

同理，考虑噪声负类标记的后验概率，有：

$$Pr[\tilde{y}=-1|x]=(1-\rho_{-1}-\rho_{+1})Pr[y=-1|x]+\rho_{+1}.$$

结合噪声正类标记的后验概率结果，可得式(5)成立.

证毕.

引理 1 证明了标记带噪数据上的加权经验风险在期望意义上等于分类器在潜在无噪声数据上的真实风险. 因此，研究者可以通过对权重进行有效估计，设计对标记噪声具备容忍性的学习算法.

在通过样本加权处理标记噪声后，本文进一步考虑复用旧特征空间的标记数据，以辅助构建在新特征空间上具有泛化能力的分类器. 在进行针对特征空间演变环境的学习理论分析之前，本文对学习算法的损失函数进行一定的假设. 本文考虑损失函数ℓ是非负凸函数，并且满足以下类似利普希茨的平滑条件.

定义 1. 损失函数的σ-admissible 性质. 如果存在$\sigma\in\mathbb{R}_+$使得对于任意 2 个分类器$g,g'\in\mathcal{G}$和对于所有样本$(x,y)\in\mathcal{X}\times\mathcal{Y}$有式(10)成立，则称损失函数$\ell$对假设空间$\mathcal{G}$是$\sigma$-admissible 的.

$$|\ell(g(x),y)-\ell(g'(x),y)|\leqslant\sigma|g(x)-g'(x)|. \tag{10}$$

值得注意的是，式(10)中定义的具有σ-admissible 性质的损失函数是较为常见的，包括平方损失、二次损失和许多其他损失函数. 具体而言，满足σ-admissible 性质的损失函数具有条件：对于$M\in\mathbb{R}_+$：$\forall g\in\mathcal{G}$和$\forall x\in\mathcal{X}$，有$|g(x)|\leqslant M$成立，且$\forall y\in\mathcal{Y}$，有$|y|\leqslant M$成立. 式(11)证明有界的最小二乘损失函数是满足条件的，其他损失函数可以通过类似的证明技术来证明. 对于$(x,y)\in\mathcal{X}\times\mathcal{Y}$和$g,g'\in\mathcal{G}$，有不等式(11)成立：

$$\begin{aligned}
&|L_2(g(x),y)-L_2(g'(x),y)|=\\
&|(g(x)-y)^2-(g'(x)-y)^2|=\\
&\left|\left[(g(x)-y)+(g'(x)-y)\right]\left[g(x)-g'(x)\right]\right|\leqslant\\
&(|g(x)-y|+|g'(x)-y|)|g(x)-g'(x)|\leqslant\\
&2\sqrt{M}|g(x)-g'(x)|.
\end{aligned} \tag{11}$$

其中式(11)的成立是因为损失函数的M有界性质. 至此，本文证明了有界最小二乘损失函数是σ-admissible 的，其中$\sigma=2\sqrt{M}$.

本文基于损失函数的σ-admissible 性质，通过构建数据差异的方式，利用旧特征空间上的标记数据辅助构建新特征空间上的分类器. 与面向特征演变环境的机器学习的相关前沿研究相似，本文假设数据流中存在数据演变阶段，这一阶段的数据具有新旧 2 个特征空间的表示. 基于旧特征空间上的标记带噪数据、新特征空间上的待预测数据，以及演变阶段数据，本文提出了容忍标记噪声的演变差异，衡量了特征演变环境中 2 个异质特征空间数据分布之间的差异.

定义 2. 容忍标记噪声的演变差异. 令权重α如式(4)定义所示，对于特征演变环境中的标记带噪数据流，存在演变阶段数据的 2 个连续数据块$\tilde{L}_t^{\rho}$与U_{t+1}，它们之间的演变差异定义为

$$\begin{aligned}
&disc_{\mathrm{NE}}\left(g,g';\mathcal{D}_t^{\rho},\mathcal{D}_{t+1}\right)=\sigma\sum_{i=1}^{k}\left|g'\left(x_i^t\right)-g\left(x_i^{t+1}\right)\right|+\\
&\sup_{f_{t+1}\in\mathcal{F}}\left|\hat{R}_{\mathcal{D}'_{t+1}}(g;f_{t+1})-\hat{R}_{\mathcal{D}_{t+1}}(g;f_{t+1})\right|,
\end{aligned} \tag{12}$$

其中$g'\in\mathcal{G}_t$，$g\in\mathcal{G}_{t+1}$是对应的 2 个分类器，$\mathcal{D}_t^{\rho}$对应着旧特征空间上标记带噪数据块L_t^{ρ}的数据分布.

容忍标记噪声的演变差异衡量了标记带噪环境中新旧 2 个不同特征空间的差异，其中关键要素在于标记存在噪声，且数据的特征空间是不同的. 定义 2 中，式(12)中等号右侧第 1 项通过演变阶段的对齐

数据对齐了分类器$g'\in\mathcal{G}_t$和$g\in\mathcal{G}_{t+1}$，右侧第2项则通过分类器衡量了演变阶段数据与新特征空间上数据的差异. 直观地说，容忍标记噪声的演变差异通过演变阶段数据建立了标记带噪环境中2个不同特征空间数据块的联系.

可以看到，本文提出的容忍标记噪声的演变差异将文献[32]所提的$\mathcal{Y}$散度差异度量的概念推广到了特征空间演变、标记存在噪声的场景，并在特征不发生演变、标记无噪声时恢复成$\mathcal{Y}$差异度量. 值得注意的是，本文所提差异度量没有使用演变阶段数据的标记，因此具有对演变阶段标记噪声的容忍度.

基于所提容忍标记噪声的演变差异，本文复用了旧特征空间的标记带噪数据，分析了分类器在新特征空间上的泛化误差上界. 可以证明，通过本文定义的容忍标记噪声的演变差异，学习者可以直接约束分类器在新特征空间上的泛化误差.

定理1. 泛化误差界. 假设损失函数ℓ是L利普希茨连续和σ-admissible的. 对于任何$\delta>0$，至少以$1-\delta$的概率，有：

$$R_{\mathcal{D}_{t+1}}(g)\leqslant\hat{R}_{\mathcal{D}_t^{\rho}}(g')+disc_{\mathrm{NE}}\left(g,g';\mathcal{D}_t^{\rho},\mathcal{D}_{t+1}\right)+O\left(1/\sqrt{n_t}+1/\sqrt{k}+1/\sqrt{n_{t+1}}\right),\tag{13}$$

其中$\hat{R}_{D_t^{\rho}}(g')$表示分类器在标记带噪数据上的加权经验损失，$disc_{\mathrm{NE}}\left(g,g';\mathcal{D}_t^{\rho},\mathcal{D}_{t+1}\right)$表示第$t$轮数据和第$t+1$轮数据之间的容忍标记噪声的演变差异.

证明：对分类器在新特征空间上的期望风险进行考察，有不等式成立：

$$R_{\mathcal{D}_{t+1}}(g)\leqslant\hat{R}_{\mathcal{D}_{t+1}}(g)+2L\mathfrak{R}_n(\mathcal{G}_t)+M\sqrt{\frac{\log(1/\delta)}{2n_{t+1}}},\tag{14}$$

$$\begin{aligned}R_{\mathcal{D}_{t+1}}(g)=&\hat{R}_{\mathcal{D}_{t+1}}(g)+\hat{R}_{\mathcal{D}_t'}(g')-\hat{R}_{\mathcal{D}_t'}(g')+\hat{R}_{\mathcal{D}_{t+1}'}(g)-\\&\hat{R}_{\mathcal{D}_{t+1}'}(g)+2L\mathfrak{R}_n(\mathcal{G})+M\sqrt{\frac{\log(1/\delta)}{2n_{t+1}}}\leqslant\\&R_{\mathcal{D}_t'}(g')+\left|R_{\mathcal{D}_t'}(g')-\hat{R}_{\mathcal{D}_t'}(g')\right|+\\&\left|\hat{R}_{\mathcal{D}_t'}(g')-\hat{R}_{\mathcal{D}_{t+1}'}(g)\right|+\\&\sup_{f_{t+1}\in\mathcal{F}}\left|\hat{R}_{\mathcal{D}_{t+1}'}(g;f_{t+1})-\hat{R}_{\mathcal{D}_{t+1}}(g;f_{t+1})\right|+\\&2L\mathfrak{R}_n(\mathcal{G})+M\sqrt{\frac{\log(1/\delta)}{2n_{t+1}}},\end{aligned}\tag{15}$$

$$\begin{aligned}R_{\mathcal{D}_{t+1}}(g)\leqslant&R_{\mathcal{D}_t'}(g')+\left|\hat{R}_{\mathcal{D}_t'}(g')-\hat{R}_{\mathcal{D}_{t+1}'}(g)\right|+\\&\sup_{f_{t+1}\in\mathcal{F}}\left|\hat{R}_{\mathcal{D}_{t+1}'}(g;f_{t+1})-\hat{R}_{\mathcal{D}_{t+1}}(g;f_{t+1})\right|+\\&2L\mathfrak{R}_n(\mathcal{G})+M\sqrt{\frac{\log(1/\delta)}{2n}}+\\&2L\mathfrak{R}_n(\mathcal{G}')+M\sqrt{\frac{\log(1/\delta)}{2k}},\end{aligned}\tag{16}$$

其中，式(14)是基于Rademacher复杂度分析的泛化误差上界，式(15)对新特征空间上的目标函数取了最大值. 式(16)同样是基于Rademacher复杂度分析的泛化误差上界. 式(16)由3项内容构成：$R_{\mathcal{D}_t'}(g')$表示分类器在t轮的演变阶段数据上的真实期望损失；$\left|\hat{R}_{\mathcal{D}_t'}(g')-\hat{R}_{\mathcal{D}_{t+1}'}(g)\right|$通过演变阶段数据建立了第$t$轮旧特征空间上标记带噪数据与第$t+1$轮新特征空间上待预测无标记数据之间的联系；$\sup_{f_{t+1}\in\mathcal{F}}\left|\hat{R}_{\mathcal{D}_{t+1}'}(g;f_{t+1})-\hat{R}_{\mathcal{D}_{t+1}}(g;f_{t+1})\right|$衡量了分类器在第$t+1$轮特征空间数据上，演变阶段数据与新特征空间所有数据之间的差异.

这3项可以依次被约束. 对于第1项，根据引理1，分类器在标记带噪数据上的加权期望损失等于分类器在潜在无噪声数据上的真实期望损失，因此可以将演变阶段数据的真实损失通过第t轮的标记带噪数据的加权经验损失和复杂度项进行上界. 因此，有式(17)成立：

$$R_{\mathcal{D}_t'}(g')=R_{\mathcal{D}_t^{\rho}}(g')\leqslant\hat{R}_{\mathcal{D}_t^{\rho}}(g')+2L\mathfrak{R}_n(\mathcal{G}_t)+M\sqrt{\frac{\log(1/\delta)}{2n_t}}.\tag{17}$$

对于式(16)不等号右侧中的第2项，算法利用损失函数的σ-admissible性质，对齐不同特征空间上的分类器g'和g. 根据损失函数的σ-admissible性质，有式(18)成立：

$$\begin{aligned}&\left|\hat{R}_{\mathcal{D}_t'}(g')-\hat{R}_{\mathcal{D}_{t+1}'}(g)\right|=\\&\left|\sum_{i=1}^{k}\ell\left(g'\left(x_i^t\right),y_i^t\right)-\sum_{i=1}^{k}\ell\left(g\left(x_i^{t+1}\right),y_i^{t+1}\right)\right|\leqslant\\&\sigma\sum_{i=1}^{k}\left|g'\left(x_i^t\right)-g\left(x_i^{t+1}\right)\right|.\end{aligned}\tag{18}$$

结合式(16)~(18)，可以得证：

$$\begin{aligned}&R_{\mathcal{D}_{t+1}}(g)\leqslant\hat{R}_{\mathcal{D}_t^{\rho}}(g')+\sigma\sum_{i=1}^{k}\left|g'\left(x_i^t\right)-g\left(x_i^{t+1}\right)\right|+\\&\sup_{f_{t+1}\in\mathcal{F}}\left|\hat{R}_{\mathcal{D}_{t+1}'}(g;f_{t+1})-\hat{R}_{\mathcal{D}_{t+1}}(g;f_{t+1})\right|+\\&O\left(1/\sqrt{n_t}+1/\sqrt{k}+1/\sqrt{n_{t+1}}\right)\leqslant\hat{R}_{\mathcal{D}_t^{\rho}}(g')+\\&disc_{\mathrm{NE}}\left(g,g';\mathcal{D}_t^{\rho},\mathcal{D}_{t+1}\right)+O\left(1/\sqrt{n_t}+1/\sqrt{k}+1/\sqrt{n_{t+1}}\right).\end{aligned}$$

证毕.

定理1证明了第$t+1$轮新特征空间上分类器的泛化误差可以被旧特征空间上标记带噪数据的加权经验风险及本文所提容忍标记噪声的演变差异所约束. 本文所提的容忍标记噪声的演变差异通过样本加权的方式对抗了标记噪声，通过在演变阶段进行分类器对齐的方式桥接了不同的特征空间. 如果样本标记是准确的，则可将α设置为1.通过对标记带噪

数据进行重新加权，本文提出的算法可以获得对潜在真实损失的无偏估计(见引理 1). 相比之下，启发式的对抗标记噪声的方法，如梯度裁剪[33]和丢弃大损失值样本[34]等，难以具备这样的能力，因此无法约束分类器在新特征空间上的期望损失的上界.

在定理 1 中，收敛率$O\left(1/\sqrt{n_t}+1/\sqrt{k}+1/\sqrt{n_{t+1}}\right)$显示出本文所提算法的收敛率与第$t$轮数据样本量、第$t+1$轮数据样本量，以及 2 轮之间演变阶段的样本量有关. 这 3 项数据样本量越大，算法对最优分类器的接近程度越高. 定理 1 建立了标记存在噪声的特征空间演变数据流中 2 个连续批次数据之间的差异度量，为分类器的泛化能力提供了理论刻画.

在证明定理 1 时，损失函数的σ-admissible 性质至关重要. 如定义 1 及式(11)中的分析所述，这一性质十分常见，适用于二次损失函数或者具有有界假设空间的损失函数. 通过利用损失函数的σ-admissible 性质，算法可以利用演变阶段的数据，将异质特征空间上的分类器g'和g进行对齐. 在对齐分类器之后，即使标记带噪数据来自不同的特征空间，算法也可以复用旧特征空间的历史数据对当前数据进行学习.

3.2 算法设计

通过本文所提容忍标记噪声的演变差异，本文分析了分类器在新特征空间上的泛化误差上界. 这一理论分析启发了本文的算法设计. 本节将提出对应的算法，直接优化分类器在新特征空间上的泛化误差，构建具有良好泛化性能的分类器. 算法首先对权重α进行估计，用以处理标记噪声；然后，算法将最小化标记带噪数据上的加权经验风险和容忍标记噪声的演变差异用以建立具有泛化能力的分类器.

基于本文所提演变差异和定理 1 中的泛化误差界，本文利用深度神经网络的特征提取能力，设计了容忍标记噪声的演变差异最小化算法 LREDM. 本文主要关注特征空间演变的 2 个连续数据块的学习问题，因此下文仅详细叙述在特征演变发生时对应的学习方案.

针对标记噪声，算法估计权重α后进行加权经验风险最小化. 根据式(4)中权重的定义

$$\alpha_i=\frac{Pr_{\mathcal{D}^\rho}(\tilde{y}_i|x_i)-\rho_{-\tilde{y}_i}}{\left(1-\rho_{\tilde{y}_i}-\rho_{-\tilde{y}_i}\right)Pr_{\mathcal{D}^\rho}(\tilde{y}_i|x_i)},$$

算法需要估计条件概率$Pr_{\mathcal{D}^\rho}(\tilde{y}_i|x_i)$和噪声率$\rho_{\pm 1}$. 本文采用密度比估计(density ratio estimation, DRE)的方法来估计条件概率. 密度比估计是一种显著减少核密度估计维度灾难的方法，它可以精确地估计高维变量的密度比. 常用的 3 种密度比估计方法包括矩匹配法、概率分类法和比值匹配法. 由于概率分类方法可能引入较大的近似误差，因此研究者在实际算法应用中更多地采用矩匹配或者比值匹配方法[35]. 其中，密度通常通过线性或非线性函数进行建模. 如果选择的再生核希尔伯特空间比较适当，那么矩匹配和比值匹配方法的近似误差可以很小. 这些方法在实践中被广泛证明了有效性[36-37].

对于条件概率$Pr_{\mathcal{D}^\rho}(\tilde{y}_i|x_i)$，本文采用比值匹配法中被广泛应用的 KLIEP 法[30]来估计. 给定样本和对应的带噪标记，KLIEP 返回其条件概率$Pr_{\mathcal{D}^\rho}(\tilde{y}_i|x_i)$. 在获得样本的条件概率后，依照文献 [26] 中的结论，噪声率满足：

$$\rho_{\tilde{y}}\leqslant Pr_{\mathcal{D}^\rho}(-\tilde{y}|x).$$

因此，本文通过对应类别在标记带噪数据上取最小条件概率的方法

$$\rho_{+1}=\min_{x_i}Pr_{\mathcal{D}^\rho}(-1|x_i),$$

$$\rho_{-1}=\min_{x_i}Pr_{\mathcal{D}^\rho}(+1|x_i),$$

来估计噪声率$\rho_{\pm 1}$.

在对权重α进行估计后，算法最小化分类器在标记带噪数据上的加权经验风险和容忍标记噪声的演变差异. 定理 1 中的泛化误差分析对应优化目标

$$\min_{g'\in\mathcal{G}_t,g\in\mathcal{G}_{t+1}}\sup_{f\in\mathcal{G}_{t+1}}\hat{R}_{\mathcal{D}_t^\rho}(g')+disc_{\mathrm{NE}}\left(g,g';\mathcal{D}_t^\rho,\mathcal{D}_{t+1}\right),\tag{19}$$

其中最后一项$disc_{\mathrm{NE}}\left(g,g';\mathcal{D}_t^\rho,\mathcal{D}_{t+1}\right)$是容忍标记噪声的演变差异，具体定义为

$$disc_{\mathrm{NE}}\left(g,g';\mathcal{D}_t^\rho,\mathcal{D}_{t+1}\right)=\left|\hat{R}_{\mathcal{D}_t'}(g')-\hat{R}_{\mathcal{D}'_{t+1}}(g)\right|+\sup_{f_{t+1}\in\mathcal{G}_{t+1}}\left|\hat{R}_{\mathcal{D}'_{t+1}}(g;f_{t+1})-\hat{R}_{\mathcal{D}_{t+1}}(g;f_{t+1})\right|.\tag{20}$$

观察式(19)和式(20)可以发现，优化问题可以看作是一个极小极大优化问题，其中最小化分类器$g,g'\in\mathcal{G}$最小化泛化误差，而最大化分类器$f_{t+1}\in\mathcal{G}_{t+1}$搜索演变差异的最坏情况. 这类学习问题在基于对抗生成的迁移学习中被广泛研究，具有成熟的优化算法[38-39]. 因此，本文通过对抗网络 DANN[38]来优化分类器的加权损失和容忍标记噪声的演变差异. 具体而言，算法交替优化式(21)和式(22)，直到损失收敛.

$$\min_{g'\in\mathcal{G}_t,g\in\mathcal{G}_{t+1}}\hat{R}_{\mathcal{D}_t^\rho}(g')+\sigma\sum_{i=1}^{k}\left|g'\left(x_i^t\right)-g\left(x_i^{t+1}\right)\right|+\hat{R}_{\mathcal{D}_{t+1}}(g;f_{t+1})-\hat{R}_{\mathcal{D}'_{t+1}}(g;f_{t+1}),\tag{21}$$

$$\sup_{f_{t+1}\in\mathcal{G}_{t+1}}\hat{R}_{\mathcal{D}_{t+1}}(g;f_{t+1})-\hat{R}_{\mathcal{D}'_{t+1}}(g;f_{t+1}).\tag{22}$$

通过最小化旧特征空间上标记带噪数据的加权经验风险以及容忍标记噪声的演变差异，算法可以

在新特征空间数据上获得一个具有良好泛化能力的分类器. 由于权重α的优化问题和分类器g'和g的优化问题不是联合凸的, 因此算法首先求解权重α来缓解标记噪声问题, 然后优化极小极大优化问题.

容忍标记噪声的演变差异最小化算法的流程如算法 1 所示.

算法 1. 容忍标记噪声的演变差异最小化算法.

输入: 每轮数据L_t^{ρ}和下一轮待预测数据U_t;

输出: 分类器$g \in \mathcal{G}_{t+1}$.

①遍历$t = 1, 2, \cdots, T$;

② 通过 KLIEP 算法估计式(4)中权重α;

③ 收到无标记数据U_{t+1};

④ 通过 DANN 算法优化式(21)和式(22)的极小极大优化问题;

⑤ 获得g_{t+1}并预测U_{t+1};

⑥ 收到标记带噪数据并存储$\tilde{L}_{t+1}^{\rho}$;

⑦结束.

4 实验与结果

本节测试了所提算法 LREDM 在现实应用场景中的性能表现. 本节实验在合成数据集和多个现实应用数据集上将所提算法与前沿的基准算法进行了对比. 具体而言, 本节实验旨在回答 3 个问题:

1)合成数据分析. 所提算法是否在数据标记带噪的情况下, 在特征空间演变的环境中有效地建立了 2 个异质特征空间的差异度量.

2)基准算法对比. 所提算法在各类现实场景中的数据集上的表现是否优于其他前沿的基准算法.

3)消融实验研究. 所提算法的每个模块是否都能够对算法性能带来提升.

本节实验采用基准数据集, 并通过对其进行修改, 模拟特征演变环境的标记带噪数据流. 为了满足实验数据对于动态特征空间的要求, 本节实验将所使用的数据集原始特征空间进行划分. 在每个时刻, 实验使用基准数据集中不同的特征空间, 并在相邻时间段中插入少量全特征空间数据, 以模拟演变阶段数据, 在 2 个特征空间上都具有特征表示. 针对类别相关的标记噪声, 本节实验考虑二分类问题, 对于某一类样本, 以一定概率将其标记翻转为另一个类别. 需要注意的是, 2 个类别的标记翻转概率可以不同. 为了测试算法的有效性, 本节实验通过随机采样的方式和通过基准数据集生成不同特征演变的标记带噪数据流, 并在不同数据流上测试算法的平均表现结果. 本节对每个数据集进行随机采样, 生成 10 条数据流, 并基于这 10 次结果的平均值和标准差来报告最终的实验结果. 针对每条数据流, 本节考虑发生 1 次特征空间演变, 并在新特征空间上测试算法的实验结果.

本文在包括 RFID, Amazon, Reuters 数据集以及其中包含的子数据集的公开数据集上进行试验, 这些数据集分布在多个现实应用领域.

1)RFID 数据集[3]. 该数据集由传感器收集的物体 RFID 数据(特征)和物体的实际位置(标记)组成. 由于物体不断到达, 传感器收集的 RFID 数据形成了数据流. 在传感器达到使用寿命之前, 学习者会在旧传感器附近放置新的传感器, 因此数据流的特征空间发生了演变. 在本节实验中, 位置索引被分为 2 类以生成相应的标记. 对于旧特征空间数据标记, 本节实验将其标记以 10% 的概率进行均匀翻转以模拟标记噪声. 根据数据的时间戳信息, 数据流的前 40% 是旧特征空间数据, 接下来的 20% 是演变阶段数据, 最后 40% 是新特征空间数据.

2) Amazon 数据集[40]. 该数据集包含用户在 2006—2008 年对亚马逊上产品的评价和对应用户的质量评分. 本节实验采用了该数据集中的 3 个子数据集, 分别是 Books, Movies 和 CDs.学习者希望根据用户的评价(特征)判断用户的质量(标记). 用户不断使用这个平台, 形成了数据流. 随着时间的推移, 旧的产品下架, 新的产品上架, 因此数据流的特征空间发生了演变. 本节实验将部分活跃用户对历史商品和当前商品的评分作为特征空间演变数据流中的数据演变阶段. 实验根据用户评价的质量将用户分为 2 类, 从而生成了一个二分类任务, 并添加了 10% 的均匀标记噪声. 根据数据的时间戳信息, 数据流的前 40% 为旧特征空间数据, 接下来的 20% 是演变阶段数据, 最后 40% 是新特征空间数据.

3) Reuters 多语言数据集[41]. 该数据集包含 5 种语言的约一万篇文档. 由于每种语言特征表示不同, 因此同一文档的不同翻译版本对应着不同的特征空间, 本节实验以此模拟特征空间演变的数据流. 本节实验将旧特征空间数据的 20% 设置为演变阶段数据, 并在新特征空间找到其对应表示. 对旧特征空间数据标记, 本节实验将文本标签分为 2 类, 并将标记以 10% 的概率均匀翻转, 用以模拟标记噪声. 所有文档都使用 TF-IDF 特征表示, 本节实验基于文档的 TF-IDF 特征进行主成分分析(PCA), 提取了文档的主要特征. 实验部分所用数据集的相关统计信息如表 2 所示.

Table 2 Statistics of the Datasets

表 2 数据集的统计信息

数据集	样本数	旧特征数	新特征数
RFID	940	78	72
Books	20 000	383	256
Movies	20 000	554	454
CDs	20 000	430	287
EN-FR	44 226	1 131	1 230
FR-SP	37 015	1 230	807
GR-IT	53 992	1 417	1 041
IT-GR	53 992	1 041	1 417

对于 LREDM 算法的实现，本节实验将采用 2 个神经网络分别作为最小化分类器和最大化分类器，每个网络都包含 5 层全连接层，并以 ReLU 函数作为激活函数. 本节实验将使用衰减学习率的在线随机梯度下降(SGD)对模型进行训练，学习率设置为 0.004，正则化权重衰减为 0.005.

4.1 合成数据分析

本节首先在合成数据上对所提算法进行分析，测试容忍标记噪声的演变差异在度量异质特征空间的标记带噪数据块上的度量性能.

在本节实验中，采用高斯分布生成数据，可以通过调整均值和方差来控制数据的差异度. 因此，可以通过对比算法对不同差异度的数据集的处理效果，来评估算法的差异度量能力. 直观上来看，算法的差异度量能力越强，就越能够准确反映数据之间的真实差异. 因此，本节实验将通过对合成数据进行分析，测试所提算法在度量异质特征空间的标记带噪数据块上的度量性能.

本节实验通过对高斯分布进行采样生成数据，合成了 3 组数据集. 每组数据集都包括 2 个不同特征空间的数据集合：一个是三维的数据集合，另一个是二维的数据集合. 为了模拟特征空间的演变，本节实验将二维数据作为旧特征空间数据，将三维数据作为新特征空间数据. 同时，二维数据和三维数据分别从不同均值的标准高斯分布中进行采样，以模拟不同差异度的学习任务.

1) 任务 1：二维数据中(旧特征空间)，正类数据从 $\mathcal{N}_x([1,1])$中生成，负类数据从 $\mathcal{N}_x([-1,-1])$中生成；三维数据中(新特征空间)，正类数据从 $\mathcal{N}_x([1,1,1])$中生成，负类数据从 $\mathcal{N}_x([-1,-1,-1])$中生成；

2) 任务 2：二维数据中(旧特征空间)，正类数据从 $\mathcal{N}_x([1,1])$中生成，负类数据从 $\mathcal{N}_x([-1,-1])$中生成；三维数据中(新特征空间)，正类数据从 $\mathcal{N}_x([1,-1,1])$中生成，负类数据从 $\mathcal{N}_x([-1,1,-1])$中生成；

3) 任务 3：二维数据中(旧特征空间)，正类数据从 $\mathcal{N}_x([1,1])$中生成，负类数据从 $\mathcal{N}_x([-1,-1])$中生成；三维数据(新特征空间)，正类从 $\mathcal{N}_x([-1,-1,-1])$中生成，负类数据从 $\mathcal{N}_x([1,1,1])$中生成；

通过计算，可以发现在三维特征空间中，从任务 1 到任务 3，同类数据的类中心距离不断增大，因此可以认为数据的差异度在一定程度上是不断增大的. 对于每个任务，本节实验生成了 1 000 个旧特征空间上的三维样本和 800 个新特征空间上的二维样本，并随机选择了 200 个样本作为演变阶段的数据.

下文将对比不同算法的数据差异度量能力. 当数据特征空间发生演变时，一个自然的做法是通过演变阶段数据建立 2 个特征空间之间的映射，以复用旧特征空间上的数据和学习模型[3]. 直观上来说，如果新旧特征空间上的数据十分相似，那么映射误差应该很小，即算法可以通过映射矩阵将新旧特征空间上的数据接近无损地转换. 为了与本文所提演变差异进行对比，下文将基于映射方法建立的新旧特征空间数据之间的重构误差称为映射误差. 记学习后的映射函数为M，映射误差可表示为：

$$err_t = |M \cdot U_t - U_{t+1}|.$$

映射误差衡量了映射函数的数据恢复能力. 假设存在一个完美的映射函数，它可以将新旧特征空间之间的表示一一映射，那么通过这个映射函数，算法可以完美恢复出新特征空间数据在旧特征空间上的表示，即映射误差为零. 与演变差异的度量方式类似，当数据越相似时，映射误差越小. 对于本文提出的基于演变差异最小化的算法，本文直接采用演变差异$disc_{\mathrm{NE}}\left(g, g'; \mathcal{D}_t^p, \mathcal{D}_{t+1}\right)$，作为衡量指标，度量所提算法所构建的新旧特征空间之间的数据差异.

本节针对 3 个合成任务测试了 LREDM 算法和 FESL 算法的性能. 由于数据存在噪声，本文统一采用 IW 算法对标记带噪数据进行加权处理后提交对比算法. 由于这 2 个衡量方式的量纲不同，无法直接比较，因此本节将所有结果进行归一化后，通过图 2 展示它们在这 3 个任务上的差异.

总体而言，本文提出的 LREDM 算法表现更加准确. 具体来说，任务 1 是 3 个学习任务中最简单的，因为新特征空间数据只是旧特征空间数据的边际分布. 图 2 表明任务 1 的演变差异明显小于任务 2 和任务 3，这符合直觉. 相比之下，FESL 算法在任务 1 中报告了最大的映射误差. 对合成数据的分析表明，LREDM 的演变差异能够在数据标记存在噪声的情况下恢复

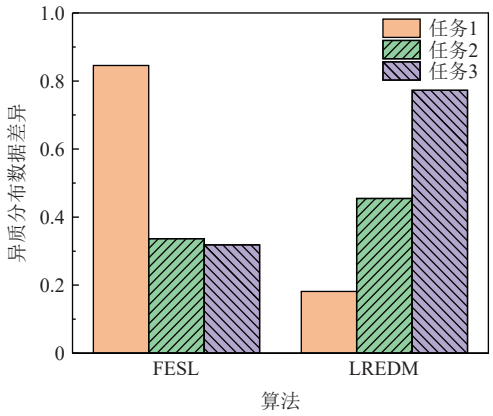


Fig. 2 Date relative discrepancy exhibited by FESL and LREDM

图 2 FESL 和 LREDM 展示的数据相对差异

异质特征空间数据块之间的差异,并为特征空间演变环境下的标记带噪数据学习问题提供了更准确的差异度量方法.

4.2 基准算法对比

本节对所提的 LREDM 算法在现实应用中的表现进行了测试,主要考虑标记带噪数据和特征空间演变的情况. 与此同时,本节还将 LREDM 算法与前沿算法的性能进行了比较.

本节实验在 3 个数据集上模拟了 8 个标记带噪数据学习任务,并与 6 种前沿基准算法进行了比较. 这些任务都涉及特征空间的演变. 其中, FESL, EDM 和 OLVF 算法是在特征空间演变环境下的前沿学习算法,能够有效处理特征空间变化下的学习任务. 然而,这些算法并未考虑标记噪声对学习的负面影响. 因此,本节实验添加了额外的标记噪声校正机制,以进行公平比较. 具体而言,本文采用重要性加权(Importance weighting, IW)机制和使用高置信度(high confidence, HC)样本机制对基于映射的代表性算法 FESL 算法进行了调整,并将这 2 种算法分别命名为 FESL+IW 和 FESL+HC. 对于 FESL+IW 算法,本节实验首先对旧特征空间的标记带噪样本进行重加权,然后使用 FESL 算法进行学习;对于 FESL+HC 算法,本节实验首先在旧特征空间上训练一个分类器,清除置信度低于阈值的标记样本,然后使用 FESL 算法进行学习. 除了这 2 种算法外,本节实验还引入了另一个基准算法 IWTS 进行两阶段学习. 与 LREDM 算法相同, IWTS 算法采用了分块学习预测的方法,并直接结合了特征演变数据学习和标记带噪数据学习的方法作为本文研究问题的基准算法. IWTS 算法首先使用 IW 在旧特征空间数据上训练模型,然后为演变阶段数据赋予伪标记,基于演变阶段的伪标记数据训练第 2 阶段模型,并对新特征空间样本进行预测.

表 3 中展示了 LREDM 算法与对比算法在各个实际应用数据集上的平均准确率比较. 实验结果表明, LREDM 算法在 8 个学习任务中均取得了最高的平均准确率,这表明 LREDM 算法是在特征空间演变环境中针对标记带噪数据学习任务的一种可行方案. 相比于基于映射的算法, LREDM 算法具有更好的性能,因为它更准确刻画了异质特征空间中标记带噪数据块之间的差异,并且直接最小化期望风险的上界. 此外,实验结果还表明, LREDM 算法总是优于 TSIW 算法. 这是因为在现实世界场景中,演变阶段的数据通常较少,很难学到准确的映射函数.

Table 3 Average Accuracy and Standard Deviation of Seven Algorithms on Different Datasets

表 3 7 种算法在不同数据集上的平均准确率与标准差 %

算法	数据集			
	RFID	Books	Movies	CDs
FESL	75.45±2.1	69.87±2.8	66.59±2.3	60.10±2.5
EDM	77.28±1.3	68.52±1.6	67.24±1.4	60.88±1.7
OLVF	80.16±2.1	71.18±2.5	67.32±1.9	59.25±1.4
FESL+IW	88.97±2.2	72.84±1.9	66.43±1.5	59.87±2.0
FESL+HC	82.55±1.9	74.37±3.1	68.86±1.5	62.78±1.2
IWTS	90.19±1.4	72.55±1.9	71.62±1.1	62.33±1.8
LREDM	92.97±1.4	76.64±3.0	75.83±1.2	68.07±2.2

算法	数据集			
	EN-FR	FR-SP	GR-IT	IT-GR
FESL	77.29±1.3	72.13±1.8	73.83±1.7	76.52±1.1
EDM	78.24±1.5	73.16±0.9	72.81±2.1	75.32±1.4
OLVF	75.89±1.6	70.60±1.7	71.52±1.8	72.36±1.2
FESL+IW	78.88±1.8	72.18±2.3	75.98±2.4	76.43±2.1
FESL+HC	77.51±1.3	72.95±1.3	75.01±1.9	77.73±1.3
IWTS	83.12±2.0	77.69±1.8	80.09±2.6	81.21±1.5
LREDM	85.70±1.3	78.92±1.6	84.87±2.1	84.01±2.9

注: "±"前后的数据分别是平均准确率和标准差.

本节实验还在语言数据集上进行了实验,测试了 LREDM 算法在文本分类场景中的性能. 表 3 中 EN-FR, FR-SP, GR-IT, IT-GR 这 4 个数据集展示了 LREDM 算法与前沿对比算法在文本数据模拟的学习任务上的实验结果. 实验结果显示,在 4 个文本分类任务中, LREDM 算法取得了最高的平均准确性,验证了该算法在文本类现实应用中的有效性.

本节实验同样测试了 LREDM 算法在不同噪声

率下的有效性. 该实验在 Movies 数据集和 EN-FR 数据集上进行, 并采用不同的噪声率来评估 LREDM 算法的性能. 表 4 展示了 LREDM 算法和 IWTS 算法在 4 种不同噪声率下的平均准确率和标准差. 例如, “0.3/0.1”表示将正类数据以 30% 的概率错误地标记为负类, 同时将负类数据以 10% 的概率错误地标记为正类. IWTS 算法作为对比算法在基准对比实验中表现出色, 因此也在表 4 中列出. 如表 4 所示, LREDM 算法在不同噪声率下均取得了更好的实验效果, 验证了 LREDM 算法在不同噪声率环境下的鲁棒性.

4.3 消融实验

本节实验在 Reuters 多语言数据集上对 LREDM 算法的每个组成部分进行了有效性测试. LREDM 算法中的关键模块包括对标记带噪数据权重的估计(即噪声率估计)、对标记带噪数据上的加权经验风险和容忍标记噪声的演变差异的最小化. 其中, 与权重估计相关的模块是 IW 算法, 没有加入 IW 算法的对比方案将标记带噪数据的标记视为准确的, 并直接进行优化. 在消融实验中, 本文所提算法被记作 LREDM(加权); 而不加入 IW 模块(不对标记噪声进行特殊处理)的对比方案被记作 LREDM(不加权).

Table 4 Average Accuracy and Standard Deviation of Two Algorithm on Datasets with Varying Noise Rates

表 4 2 种算法在 2 种数据集上不同噪声率下的平均准确率与标准差 %

算法	Movies 数据集的不同噪声率			
	0.1/0.1	0.2/0.1	0.3/0.1	0.2/0.2
IWTS	71.62±1.1	70.52±0.9	68.63±1.5	69.40±1.7
LREDM	75.83±1.2	74.86±1.3	73.58±1.4	73.84±1.6
算法	EN-FR 数据集的不同噪声率			
	0.1/0.1	0.2/0.1	0.3/0.1	0.2/0.2
IWTS	83.12±2.0	81.05±2.4	78.86±2.0	80.04±2.6
LREDM	85.70±1.3	84.11±1.9	82.56±2.3	83.39±2.5

注: “±”前后的数据分别是平均准确率和标准差.

本节实验测试了 LREDM 算法在 Reuters 多语言数据集上 6 对不同语言之间进行特征空间演变的标记带噪数据学习. 图 3 展示了 LREDM 算法在这些任务中的加权经验风险、容忍标记噪声的演变差异和平均准确率随迭代次数的变化情况. 实验结果表明, 随着演变差异的减少, LREDM 算法在这 6 项任务中的平均准确率均有所提高, 这验证了最小化容忍标记噪声的演变差异对于特征空间演变环境中标记带

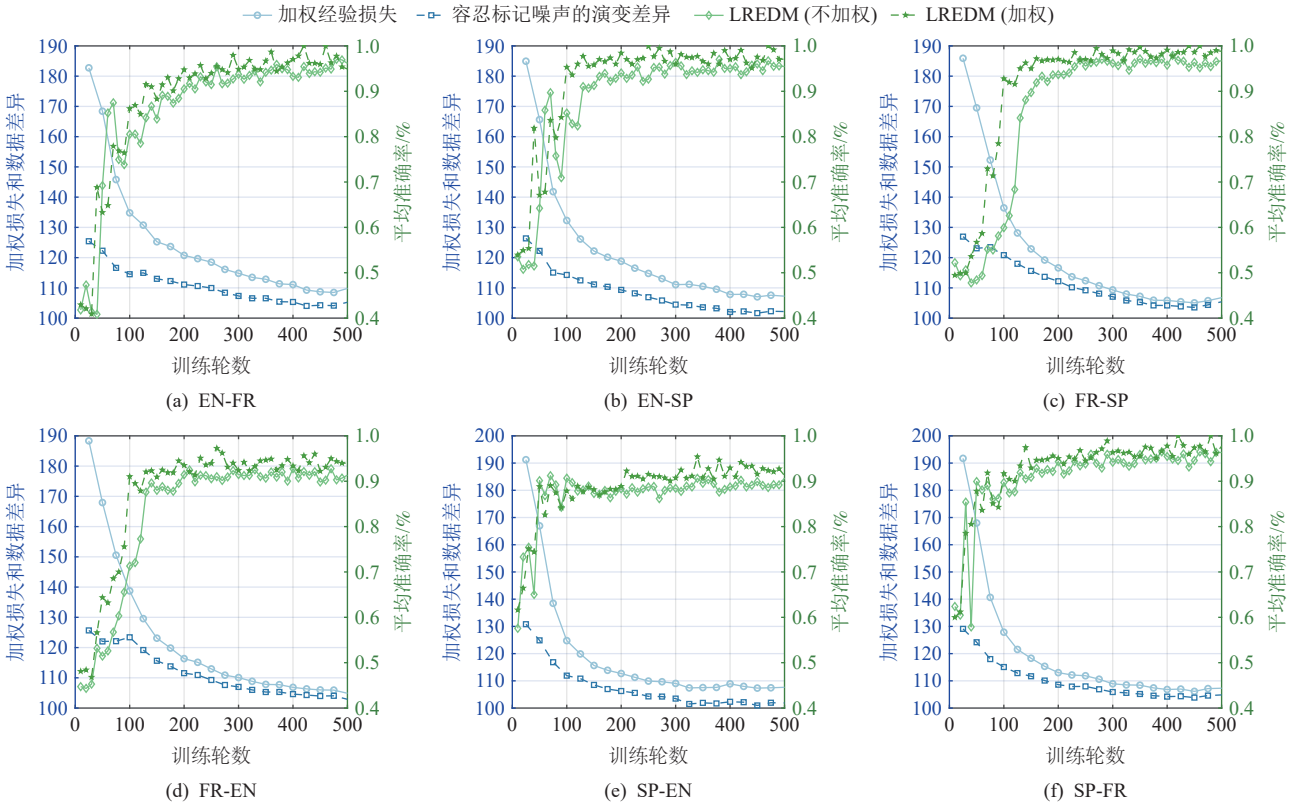


Fig. 3 Weighted empirical risk, label noise robust evolving discrepancy and average accuracy of LREDM algorithm on six groups of experiment simulated by Reuters dataset

图 3 LREDM 算法在 Reuters 数据集模拟的 6 组实验上的标记带噪数据的加权经验风险、容忍标记噪声的演变差异和平均准确率

噪数据学习问题的重要性. LREDM(加权)算法通过学习权重对标记带噪数据进行去噪,然后通过最小化容忍标记噪声的演变差异来减小异质特征空间的数据差异. 实验结果显示,与不带标记噪声进行处理的对比方案 LREDM(不加权)相比,通过应用 IW 算法学习权重的 LREDM(加权)算法获得了具备更好泛化性能的分类器. 对于存在标记噪声的学习问题,IW 算法可以缓解标记噪声带来的影响. 因此,消融实验的结果验证了本文所提的容忍标记噪声的演变差异度量的有效性,并验证了 LREDM 算法中每个组件的有效性.

因此,通过对 LREDM 算法在 Reuters 新闻分类任务上的消融实验,本节验证了 LREDM 算法中每个组件的有效性.

5 结 论

本文研究了在特征空间演变环境下的标记带噪数据学习问题,这个问题在实际应用中非常重要且普遍存在. 由于数据标记存在噪声并且特征空间可能演变,如何设计学习算法,有效利用旧特征空间中的标记带噪数据,使得分类器在新特征空间上具有良好的泛化能力,是一个极具挑战性的问题. 为了应对这一挑战,本文提出了容忍标记噪声的演变差异来建立具有不同特征空间的连续数据块的差异度量,并且分析了分类器在特征空间演变的标记带噪数据流上的泛化误差. 基于所提理论,本文设计了一种名为 LREDM 的演变差异最小化算法,并使用深度神经网络进行实现. 通过对合成数据的分析研究,本文验证了所提出的容忍标记噪声的演变差异可以建立异质特征空间的标记带噪数据块之间的差异度量. 同时,多个实际应用数据的实验表明了本文所提出的算法在性能和稳健性方面的优越性.

本文未来的工作主要集中在 2 个方面:首先,当前所提出的算法在每一轮中需要求解一个极小极大化的优化问题,这一求解过程耗时较多. 因此,如何进一步降低算法的运行时间,以更好地适应数据流学习是下一步重要的研究方向. 其次,本文考虑了与样本无关的均匀标记噪声,但现实应用中的标记噪声往往更为复杂. 因此,如何建立更加有效的求解算法以应对现实应用中复杂的标记噪声,是未来的重要研究课题.

作者贡献声明:张震宇负责提出模型,完成实验、初稿写作和论文修改;姜远负责写作指导和修改审定.

参 考 文 献

[1] Zhou Zhihua. Open-environment machine learning[J]. National Science Review, 2022, 9(8): nwac123

[2] Zhou Zhihua. A brief introduction to weakly supervised learning[J]. National Science Review, 2018, 5(1): 44−53

[3] Hou Bojian, Zhang Lijun, Zhou Zhihua. Learning with feature evolvable streams[C] //Advances in Neural Information Processing Systems 30. Cambridge, MA: MIT, 2017: 1416−1426

[4] Zhang Zhenyu, Zhao Peng, Jiang Yuan, et al. Learning with feature and distribution evolvable streams[C] //Proc of the 37th Int Conf on Machine Learning. New York: ACM, 2020: 11317−11327

[5] Cesa-Bianchi N, Dichterman E, Fischer P, et al. Sample-efficient strategies for learning in the presence of noise[J]. Journal of the ACM, 1999, 46(5): 684−719

[6] Natarajan N, Dhillon I S, Ravikumar P K, et al. Learning with noisy labels[C] //Advances in Neural Information Processing Systems 26. Cambridge, MA: MIT, 2013: 1196−1204

[7] Song H, Kim M, Lee J G. Selfie: Refurbishing unclean samples for robust deep learning[C] //Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 5907−5915

[8] Ben-David S, Blitzer J, Crammer K, et al. Analysis of representations for domain adaptation[C] //Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT, 2006: 137–144

[9] Mansour Y, Mohri M, Rostamizadeh A. Domain adaptation: Learning bounds and algorithms[C] //Proc of the 22nd Conf on Learning Theory. New York: ACM, 2009: 18–29

[10] Cortes C, Mohri M, Medina A M. Adaptation based on generalized discrepancy[J]. Journal of Machine Learning Research, 2019, 20(1): 1−30

[11] Dietterich T G. Steps Toward Robust Artificial Intelligence[J]. AI Magazine, 2017, 38(3): 3−24

[12] Zhou Zhihua. Learnware: On the future of machine learning[J]. Frontiers of Computer Science, 2016, 10(4): 589–590

[13] Guan S U, Li Shanchun. Incremental learning with respect to new incoming input attributes[J]. Neural Processing Letters, 2001, 14: 241−260

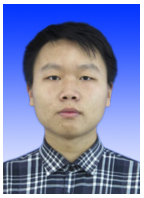
[14] Zhang Qin, Zhang Peng, Long Guodong, et al. Online learning from trapezoidal data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(10): 2709−2723

[15] Liu Yanfang, Li Wenbin, Gao Yang. Passive-aggressive learning with feature evolvable streams[J]. Journal of Computer Research and Development, 2021, 58(8): 1575−1585 (in Chinese)
(刘艳芳, 李文斌, 高阳. 基于被动-主动的特征演化流学习[J]. 计算机研究与发展, 2021, 58(8): 1575−1585)

[16] Hou Chenping, Zhou Zhihua. One-pass learning with incremental and decremental features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2776−2792

[17] Liu Zhaoqing, Gu shilin, Hou Chenping. Online classification algorithm with feature inheritably increasing and decreasing[J]. Journal of Computer Research and Development, 2022, 59(8):

1668−1682 (in Chinese)
(刘兆清, 古仕林, 侯臣平. 面向特征继承性增减的在线分类算法[J]. 计算机研究与发展, 2022, 59(8): 1668−1682)
[18] He Yi, Wu Baijun, Wu Di, et al. Online learning from capricious data streams: A generative approach[C] //Proc of the 28th Int Joint Conf on Artificial Intelligence. Macao, SAR China: Morgan Kautman, 2019: 2491–2497
[19] Beyazit E, Alagurajah J, Wu Xingdong. Online learning from data streams with varying feature spaces[C] //Proc of the 33rd AAAI Conf on Artificial Intelligence. Menlo Park: CA: AAAI, 2019: 3232−3239.
[20] Dong Jiahua, Cong Yang, Sun Gan, et al. Evolving metric learning for incremental and decremental features[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(4): 2290−2302
[21] Hou Bojian, Zhang Lijun, Zhou Zhihua. Prediction with unpredictable feature evolution[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(10): 5706−5715
[22] Angluin D, Laird P. Learning from noisy examples[J]. Machine Learning, 1988, 2: 343−370
[23] Aslam J A, Decatur S E. On the sample complexity of noise-tolerant learning[J]. Information Processing Letters, 1996, 57(4): 189−195
[24] Gao Wei, Wang Lu, Zhou Zhihua. Risk minimization in the presence of label noise[C] //Proc of the 30th AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2016: 1575−1581
[25] Arora S, Ge Rong, Moitra A. Learning topic models-going beyond SVD[C] //Proc of the 53rd IEEE Annual Symp on Foundations of Computer Science.Piscataway, NJ: IEEE, 2012: 1−10
[26] Liu Tongliang, Tao Dacheng. Classification with noisy labels by importance reweighting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(3): 447−461
[27] Zhang Zhenyu, Zhao Peng, Jiang Yuan, et al. Learning from incomplete and inaccurate supervision[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(12): 5854−5868
[28] Scott C, Blanchard G, Handy G. Classification with asymmetric label noise: Consistency and maximal denoising[C] //Proc of the 26th Conf on Learning Theory. Berlin: Springer, 2013: 489−511
[29] Ramaswamy H, Scott C, Tewari A. Mixture proportion estimation via kernel embeddings of distributions[C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 2052−2060
[30] Sugiyama M, Nakajima S, Kashima H, et al. Direct importance estimation with model selection and its application to covariate shift adaptation[C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2007: 1433–1440
[31] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test[J]. Journal of Machine Learning Research, 2012, 13(1): 723−773
[32] Mohri M, Muñoz-Medina A. New analysis and algorithm for learning with drifting distributions[C] //Proc of the 23rd Int Conf on Algorithmic Learning Theory. Berlin: Springer, 2012: 124−138
[33] Menon A K, Rawat A S, Reddi S J, et al. Can gradient clipping mitigate label noise?[C/OL] //Proc of the 8th Int Conf on Learning Representations. 2020. https://openreview.net/forum?id=rklB76EKPr
[34] Han Bo, Yao Quanming, Yu Xingrui, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels[C] //Proc of Advances in Neural Information Processing Systems 31, Cambridge, MA: MIT, 2018: 8536−8546
[35] Kanamori T, Suzuki T, Sugiyama M. Theoretical analysis of density ratio estimation[J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2010, 93(4): 787−798
[36] Huang Jiayuan, Gretton A, Borgwardt K, et al. Correcting sample selection bias by unlabeled data[C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2006: 601−608
[37] Kanamori T, Hido S, Sugiyama M. A least-squares approach to direct importance estimation[J]. Journal of Machine Learning Research, 2009, 10: 1391−1445
[38] 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
[39] Zhang Yuchen, Liu Tianle, Long Mingsheng, et al. Bridging theory and algorithm for domain adaptation[C] //Proc of the 36th Int Conf on Machine Learning. New York: ACM, 2019: 7404−7413
[40] McAuley J, Targett C, Shi Qinfeng, et al. Image-based recommendations on styles and substitutes [C] //Proc of the 38th int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2015: 43−52
[41] Amini MR, Usunier N, Goutte C. Learning from multiple partially observed views-an application to multilingual text categorization[C] //Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2009: 28−36



Zhang Zhenyu, born in 1994. PhD. His main research interests include machine learning and data mining.
张震宇, 1994 年生. 博士. 主要研究方向为机器学习与数据挖掘.

Jiang Yuan, born in 1976. PhD, professor, PhD supervisor. Her main research interests include artificial intelligence, machine learning, and data mining.
姜　远, 1976 年生. 博士, 教授, 博士生导师. 主要研究方向为人工智能、机器学习和数据挖掘.