小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2024-0577
2026 年 2 月 第 2 期
Vol.47 No.2 2026

类簇分布引导的标签噪声检测与清洗

姜高霞[1],张 妮[1],王文剑[2,3]

[1](山西大学 计算机与信息技术学院,太原 030006)
[2](山西大学 数据智能与认知计算山西省重点实验室,太原 030006)
[3](山西警察学院 网络安全保卫系,太原 030401)
E-mail:wjwang@sxu.edu.cn

摘 要: 标签噪声通常会误导分类器训练并对分类性能产生负面影响.利用聚类技术可以实现类簇分布引导的标签噪声辅助检测,但存在特征语义和标签语义不一致的问题,导致噪声识别过于敏感.为了充分挖掘数据分布信息以提升噪声检测准确度,本文提出一种类簇分布引导的噪声检测与清洗方法(Cluster Distribution-Guided Label Noise Detection and Cleaning,CDGDC).该方法通过类别编码的属性增强机制实现特征语义和标签语义的自适应融合,并构造近邻特征曲线来进一步排除伪标签噪声,最后对识别出的噪声进行纠正或过滤的针对性清洗以提升数据质量.实验结果表明,所提方法在模拟和真实含噪数据集上能够准确识别标签噪声,并有效提高了分类器泛化性能.

关 键 词: 标签噪声;类簇分布;语义一致性;噪声检测与清洗;泛化性能
中图分类号: TP181 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)02-0326-10

Cluster Distribution-guided Label Noise Detection and Cleaning

JIANG Gaoxia[1],ZHANG Ni[1],WANG Wenjian[2,3]

[1](School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China)
[2](Key Laboratory of Data Intelligence and Cognitive Computing of Shanxi Province,Shanxi University,Taiyuan 030006,China)
[3](Department of Network Security,Shanxi Police College,Taiyuan 030401,China)

Abstract: Label noise usually misleads classifier training and negatively impacts classification performance. The use of clustering technology can assist cluster distribution-guided label noise detection. However,there exists the problem of inconsistency between feature semantics and label semantics,and the noise identification is too sensitive. In order to explore the data distribution information fully and improve the accuracy of noise detection,this paper proposes a cluster distribution-guided label noise detection and cleaning(CDGDC) method. The method develops an adaptive fusion of feature semantics and label semantics through the category attribute enhancement mechanism,constructs the feature curve of the nearest neighbors to further screen out the true label noise,and finally carries out targeted cleaning for the detected noise to improve the data quality. Experimental results show that the proposed method can accurately identify the label noise and improve the model generalization performance on simulated and real-world noisy datasets.

Keywords: label noise;cluster distribution;semantic consistency;noise detection and cleaning;generalization performance

0 引 言

数据的标签质量在机器学习研究中至关重要,它会直接影响模型的泛化能力[1].目前主要通过专家标注或众包等方式获取标签信息,但是由于标注过程中信息描述不足、数据质量差、标注人员的主观性、使用的标注算法效果差等原因[2-5],不可避免地会引入一定程度的噪声使得数据质量参差不齐.在监督学习中,根据噪声的位置可以将噪声分为标签噪声和特征噪声,这些噪声会对机器学习算法的泛化能力产生一些负面影响[6].文献[7]表明相对于特征噪声,标签噪声对分类器的影响更大.为了提高模型在噪声环境下的预测能力,需要探索如何降低噪声对数据质量和模型性能的影响,以及如何在噪声环境下训练出更鲁棒的模型.

标签噪声通常采用以下两类方法处理:通过集成、加权以及损失函数重构等方式构建噪声鲁棒模型[8]和基于数据层面的标签噪声过滤或纠正[9].标签噪声的鲁棒性建模受到噪声影响并没有表现出完全的鲁棒性,即使是专门针对标签噪声的模型也会或多或少受噪声影响.从数据层面来看,主要是对噪声数据进行检测,识别出的噪声数据在模型训练前就被删除或纠正,从而降低对建模的影响.常用的噪声识别策略主要基于模型预测或离群点检测,前者是根据模型的多数预测类别或观测类别的预测置信度来识别标签噪声[10];后者按观测标签逐类识别数据集中显著偏离其它样本的数据点,来发现可能存在的噪声样本[11].

收稿日期:2024-12-04 收修改稿日期:2025-01-13 基金项目:国家自然科学基金项目(62476157,62276161,U21A20513)资助;山西省重点研发计划项目(202302010101007)资助;山西省基础研究计划面上项目(202303021221055)资助;教育部人文社科项目(24YJAZH022)资助. 作者简介:**姜高霞**,男,1987 年生,博士,副教授,博士生导师,CCF 会员,研究方向为机器学习和数据挖掘;**张 妮**,女,1998 年生,硕士研究生,研究方向为机器学习;**王文剑**(通信作者),女,1968 年生,博士,教授,博士生导师,CCF 杰出会员,研究方向为机器学习和数据挖掘.

标记数据的特征丰富多样,而标签信息少而精.在标签噪声识别中,通常关注较多的是数据的标签信息,这导致一般的标签噪声识别算法难以充分挖掘特征信息.而基于类簇分布的噪声识别方法可以较为充分地利用特征信息来评价标签质量.在特征分布和标签语义一致的前提下,同一类簇的样本应当具有相同的类标签,因此本文拟采用类簇分布思想和聚类技术重点关注类簇标记与类标签不一致的样本,辅助识别标签噪声进而提高训练数据的质量.

然而,基于类簇分布的噪声识别算法会出现以下问题:1)语义不一致问题,指特征语义和标签语义关注的类别信息不一致;2)噪声识别不准确问题,类边界附近的样本容易被错误识别为噪声;3)超参数设置问题,阈值设定不合适可能将部分干净样本误判为噪声样本.针对上述问题,本文采用特征增强方式以缓解语义不一致问题,并通过近邻特征曲线提高标签噪声识别准确率.最后提出了类簇分布引导的标签噪声检测与清洗方法(Cluster Distribution-Guided Label Noise Detection and Cleaning,CDGDC),该方法借助类簇分布信息识别标签噪声,并将识别出来的错误标签进行纠正或过滤.本文主要贡献如下:

1)设计了基于类簇分布的标签噪声识别方法,针对属性与标签语义不一致性问题,将标签信息通过自适应加权融入特征矩阵,以增强属性信息并实现属性和标签的语义统一;

2)针对噪声识别中的假噪声问题,提出近邻特征曲线用于区分真假标签噪声,以此提升标签噪声检测准确率;

3)建立了依赖任务难度的噪声清洗模式,对二分类数据集上的噪声标签进行翻转纠正,在多分类数据集上根据标签一致性进行噪声过滤或纠正处理,以提升分类器泛化能力;

4)在标准数据集和真实性别数据集上的实验结果表明,所提方法在噪声检测准确率和提升分类器泛化性能方面有明显优势.

1　相关工作

本节主要对所用的相关技术和研究进展进行简要介绍.

1.1　类簇分布

聚类假设认为数据点可以被划分为若干个自然形成的群组,同一类簇中的数据可能有相似特性,应当具有相同的某种语义类标记.在理想的数据集中,每个类簇应该有明确的分布范围,这样的分布有助于算法有效地学习和识别数据的模式,提高分类或聚类的准确性.然而在现实世界的数据集中,类簇分布往往比较复杂,如可能存在重叠的类簇、模糊的边界,甚至是孤立的异常点等,这些因素都会增加学习任务的难度,因此对算法的鲁棒性和分类器的泛化能力提出更高的要求.

聚类是一类常用的无监督数据分析方法,它无需标记就可以将数据按照相似性分成不同的类簇,这将有助于识别和过滤那些与其他数据点差异较大的类标签噪声,从而提高数据的可靠性.聚类可分为基于划分、基于层次、基于密度、基于网格、基于模型的方法[12,13].K 均值法是一种经典的基于划分的聚类方法,通过选择初始聚类中心,分配每个样本至最近的中心,重新计算中心点,直到簇不再变化;基于层次的聚类方法是生成一个簇的层次结构,一般用树状结构表示;基于密度的方法通过识别高密度区域来形成簇,能够有效发现任意形状的簇;基于网格的方法是把数据空间转化成一个网格结构,将样本特征映射到网格单元上,通过合并具有相似特性的相邻网格单元而聚类;基于模型的方法会假设数据来自某种概率分布,并通过优化分布参数来确定簇的划分.聚类之后的类簇分布分析有助于识别数据中的异常值,比如离簇中心很远或属于非常小类簇的数据点有可能是噪声.

监督学习数据集通常包含所关注任务的类标签,而聚类数据不需预先标注,聚类算法根据数据的内在的特征或属性将数据点生成若干类簇.通过对比已有观测类标签和聚类类簇标签,可以进行初步的标签噪声检测与分析,尤其是在非平衡或非对称的噪声数据集中,聚类算法的辅助检测效果可能会更显著.与之形成对比的是,基于模型预测的方法在面对非平衡或非对称噪声数据时,类边界可能会发生不同程度的偏移从而导致一些类边界附近的标签噪声识别出现错误.而利用聚类技术可能会减少这种误判情况,如层次聚类可以将大类划分为多个子类从而避免出现类边界偏移的现象,因此聚类分析和数据的类簇分布信息对于标签噪声识别和清洗具有重要意义.

1.2　标签噪声识别与清洗

标签噪声清洗方法通常指对识别出的标签噪声进行过滤或纠正处理,以此来提高数据质量[7].标签纠正如果出错可能引入额外噪声,样本过滤出错会浪费有效数据,因此标签噪声识别是数据清洗的重要前提.标签噪声过滤方法主要包括近邻过滤、分类预测过滤和集成过滤[14].

基于近邻的过滤方法中,如果一个样本与其近邻样本的标签不一致,则认为该样本含标签噪声,全近邻(All Nearest Neighbors,ANN)[15]给定多个不同的近邻 K 值,分类预测其是否为噪声标签,这些方法对近邻参数 K 的选取较敏感,参数的随机设定会影响模型鲁棒性,降低识别准确率,甚至过度清洗数据.分类过滤器(Classification Filter,CF)[16-18]利用交叉验证方法产生预测标签,当样本的预测标签和观测标签不一致时,认为该样本为噪声并进行过滤.此外,基于不同集成策略的过滤方法应用也很广泛,通过基分类器对样本进行预测,根据多数投票和一致性投票准则移除数据集中的噪声.多数投票过滤器(Majority Vote Filter,MVF)[19]是利用 KNN、C4.5 和朴素贝叶斯 3 个不同的分类器组合预测结果来识别噪声,比单个分类器过滤具有更好的正确率,它给定阈值,依据阈值和集成算法的投票结果大小关系过滤潜在的噪声样本.作为迭代方法的代表,迭代划分过滤算法(Iterative-Partitioning Filter,IPF)[20]将数据集分成几个子集分别训练模型并预测,通过多次迭代、多个模型的预测结果进行投票,判断最终分类结果,直至识别噪声数量到设定样本阈值.文献[14]提出一种完整而高效的完全随机森林方法(Completed Random Forest,CRF),通过构建完全随机树来评估样本被不同类样本包围的水平,进而确定样本噪声强度,那些大于给定的强度阈值的节点对象被识别为噪声,该过程不受不同特征权值的影响,也不受特定分类器缺点的限制.此外,所使用的投票机制使其能够有效地处理通常被特征噪声污染的高维数据集,但是阈值的设定直接会影响噪声识别的目标函数,(过大或过小)的值都会产生负面效应[21].

此外,还有部分方法将标签进行纠正来提高数据集的利用率. 噪声标签识别与纠正的置信度预测方法(Confidence Prediction method for noise label identification and Correction, CPRC)[22]考虑样本间标签误差与距离的置信度计算,使用阈值和预测标签识别与纠正噪声标签. 基于动态重采样(Dynamic Resampling Noise Correction, DRNC)[23]的新型标签噪声校正方法充分利用数据集中的有效信息多次迭代划分噪声集和干净集,分类器通过投票纠正错误标签,但如果噪声数量过高会使结果不佳. 文献[24]提出了一种概率抽样(Probabilistic Sampling, PSAM)方法,利用概率多重投票的思想赋值给干净和错误标记样本不同的置信度值,识别选中更多的干净样本,但是该过程仍然需要设定阈值区间,同时忽视了不同数据集之间的差异,容易导致结果误差大,最终影响算法的噪声识别能力.

2 基于类簇分布的标签噪声检测与清洗方法

本节设计了标签噪声识别和处理的总体思路,先通过类簇标签对数据集进行初步检测,筛选出各簇中的少数类样本作为可疑噪声样本,然后利用近邻特征曲线排除假噪声样本,从而挑选出所识别的噪声标签. 最后依据噪声识别结果对标签进行清洗,以提高数据集质量和分类器泛化性能.

2.1 标签噪声检测

通过检测数据集中的标签噪声并剔除或修正错误标签,可以为分类器提供一个更准确的数据基础和训练环境. 本文通过类簇分布思想和聚类技术初步检测标签噪声. 在特征分布和标签语义一致的前提下,同一类簇的样本应当具有相同的类标签,因此可以重点关注类簇标记与类标签不一致的样本,辅助识别标签噪声. 具体标签噪声识别过程如图 1 所示,其中原始样本属于3个类别,经过聚类后形成4个类簇(通常设定类簇数多于类别数),其中类簇1和类簇2组成类标签

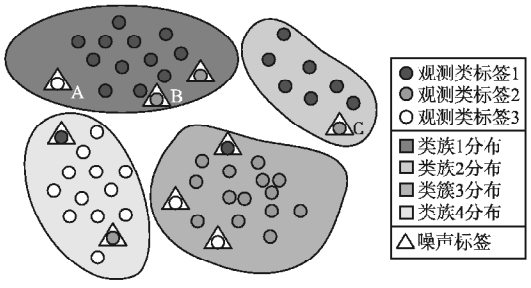


图 1 基于类簇分布的噪声识别示意图

Fig. 1 Schematic diagram of noise identification based on cluster distribution

1,类簇 3 对应类标签 2,类簇 4 对应类标签 3. 标签噪声可通过类簇的多数类标签来识别,如图 1 中类簇 1 和类簇 2 的多数样本具有类标签 1,但样本 A、样本 B 和样本 C 的标签均不是类标签 1,故可将这些样本初步认定为可疑或候选标签噪声. 另外,同类标签的噪声识别通常不受类簇划分的影响,如样本 A 和样本 C 虽然属于不同类簇(真实标签很可能都是类标签 1),但不影响将其识别为可疑标签.

在利用类簇划分识别噪声的过程中,可能出现语义不一致和噪声识别不准确的问题. 图 1 所示案例是在类簇语义和标签语义一致的前提下得到的识别结果,然而特征语义和标签语义并非总是一致的,这时类簇分布和类标签分布的对应关系可能会比较混乱,使得噪声识别变得更加困难. 此外,同一类簇中可能包含不同类样本,这时容易将类簇中的少数类样本统一识别为噪声样本,即将干净样本识别为噪声样本. 因此还需进一步甄别标签噪声的真假,以提高噪声识别准确性.

2.2 特征自适应增强

本文将标签信息融入特征矩阵以缓解语义不一致的问题. 假设 $\boldsymbol{X}$ 为原始特征矩阵(n 是样本量,m 是特征数),$\boldsymbol{Y}$ 为原始标签向量:

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, Y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$\boldsymbol{H}$ 为独热(one-hot)编码形式的标签矩阵,$\boldsymbol{H}$ 的每一列取值表示样本是否属于某一类:

$$H=(h_{ic})_{n\times C}, h_{ic}=\mathbb{I}(y_i=c)=\begin{cases}1, & y_i=c \\ 0, & y_i\neq c\end{cases} \tag{1}$$

其中 $c=1,2,\cdots,C$,C 表示观测类别数. 将原始特征矩阵和标签矩阵加权合并可以得到增强后的特征矩阵:

$$X'=[X,W\odot H]=\left[\begin{array}{cccc|cccc} x_{11} & x_{12} & \cdots & x_{1m} & w_1h_{11} & w_2h_{12} & \cdots & w_Ch_{1C} \\ x_{21} & x_{22} & \cdots & x_{2m} & w_1h_{21} & w_2h_{22} & \cdots & w_Ch_{2C} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} & w_1h_{n1} & w_2h_{n2} & \cdots & w_Ch_{nC} \end{array}\right] \tag{2}$$

其中 $W=\begin{bmatrix} w_1 & w_2 & \cdots & w_C \\ w_1 & w_2 & \cdots & w_C \\ \vdots & \vdots & \ddots & \vdots \\ w_1 & w_2 & \cdots & w_C \end{bmatrix}_{n\times C}$,各类的权重由语义一致性决定,这里的语义一致性通过模型在每个类的预测准确率来度量:

$$w_c=1-\sqrt[C]{Acc_c}, \tag{3}$$

式(3)中 $w_c(c=1,2,\cdots,C)$ 表示各个类别的标签权重,Acc_c 表示各类的预测准确率. 通常,分类精度越高,表明特征和标签的语义一致性较高,此时特征矩阵不需要太多的标签信息,故设置较小的权重;反之分类精度越低,语义一致性就越低,此时需要在增强特征矩阵中设置较大的权重来强化标签信息.

通过利用标签矩阵对原始特征矩阵进行加权增强再聚类,实现了特征和标签的自适应融合,可以缓解基于类簇分布的标签噪声识别中出现的语义不一致问题.

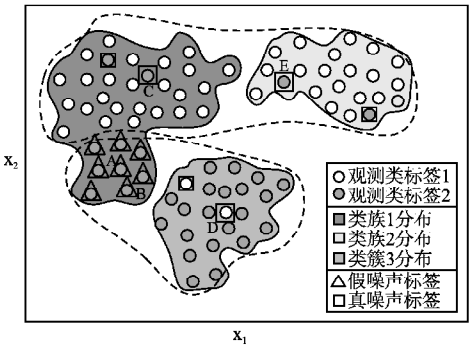
2.3 真假标签噪声识别

在 2.1 节基于类簇分布的标签噪声检测过程中,每个类簇中的少数类样本(标签不同于多数类的样本)组成可疑噪声样本集,然而并非所有少数类样本都为噪声.

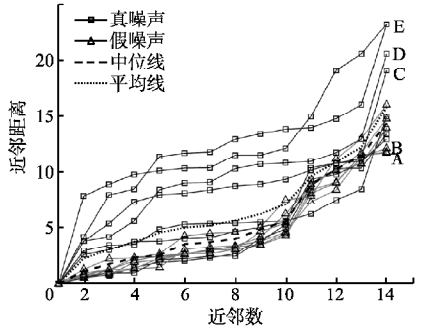
语义不一致问题可能导致聚类后出现以下两种不理想情况:1)同一大类的样本聚类后分裂为不同类簇,此时该类样

本在不同类簇中通常属于多数类,因此这种情况对噪声识别影响不大,如图 1 中类簇 1 和类簇 2 对应类标签 1;2)不同类的样本聚类之后出现在同一类簇中,此时会出现“大类吃掉小类”的情况,即将类簇中样本数相对较少的那类样本统一列为可疑噪声,如果不加区分和再识别,很容易出现过度清洗. 针对该问题本文将可疑噪声样本集 D_N 分为真噪声集 D_{TN} 和假噪声集 D_{FN},真噪声是由偶然错误标记导致的,在 D_N 中通常比较随机和分散;假噪声很可能源自语义不一致问题,一些假噪声属于“被吃掉的同一个小类”,在 D_N 中的位置相对集中. 因此可以根据可疑噪声样本的集中程度来区分真假噪声.

以二分类数据集为例,图 2(a)中数据原始分布包含两个观测类(两个黑色虚线框圈住的样本),数据聚类之后形成 3 个类簇. 其中,观测类 1 在聚类后主要分成类簇 1 和类簇 2,其中少数的错误标记样本可以被准确识别为噪声,如样本 C 和样本 E;观测类 2 在聚类后主要被分到类簇 1 和类簇 3. 由于观测类 2 被分到类簇 1 的样本数量相对较少,在类簇 1 中属于少数类样本,因此它们会被误判为噪声,如样本 A 和样本 B. 为排除 D_N 中的假噪声 D_{FN},本文统计每个可疑噪声样本的近邻分布关系,并在 D_N 上定义近邻特征曲线来区分真假噪声,此曲线由每个可疑噪声样本与其余可疑噪声样本的前 K 个近邻的距离所得,如图 2(b)所示.



(a)数据分布



(b)近邻特征曲线

图 2　二分类近邻特征曲线示意图

Fig. 2　Schematic diagram of bi-class nearest-neighbor characteristic curves

在图 2(a)中由于两类样本混入类簇 1,导致类簇 1 下方的 8 个三角形所框样本被错误识别为可疑噪声. 同时在所有可疑噪声样本中这些假噪声的位置相对更为接近,每个假噪声样本的前 8 个近邻(含样本本身)距离较小,故假噪声样本的近邻特征曲线在近邻数较少时(近邻数 < 8)比真噪声样本的曲线更靠下;而真实标签噪声位置相对更分散,每个真噪声样本的前 8 个近邻距离偏大,对应曲线更靠上方,如样本 C 和样本 D. 基于上述分析,可以设置一条参考线,认为特征曲线的左侧位于参考线下方的样本属于 D_{FN}. 由于各个样本的特征曲线变化趋势不尽一致,因此选择特定的分位曲线(采用可调节的 p 分位线作为参考线,其中每个值为所有可疑噪声样本 K 近邻距离的 p 分位数). 如果某样本近邻特征曲线的大部分距离值小于该参考线上对应位置的距离值,则认为该样本为假噪声,应当属于 D_{FN}.

具体计算过程如下:

1)计算所有可疑噪声样本 D_N 的前 K 个近邻距离矩阵 $\boldsymbol{V}$:

$$V=\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1k} & \cdots & v_{1K} \\ v_{21} & v_{22} & \cdots & v_{2k} & \cdots & v_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ v_{s1} & v_{s2} & \cdots & v_{sk} & \cdots & v_{sK} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{r1} & v_{r2} & \cdots & v_{rk} & \cdots & v_{rK} \end{bmatrix},$$

其中 r 表示 D_N 所含样本数量,v_{sk} 表示第 s 个可疑样本到其 K 近邻的距离.

近邻距离矩阵 $\boldsymbol{V}$ 可以分解为若干个行向量 V_s 也可以分解为若干个列向量 V_k^*:

$$V=\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_s \\ \vdots \\ V_r \end{bmatrix}=\begin{bmatrix} V_1^* & V_2^* & \cdots & V_k^* & \cdots & V_K^* \end{bmatrix} \tag{4}$$

其中,$V_s=[v_{s1}\quad v_{s2}\quad \cdots\quad v_{sK}]$,$V_k^*=[v_{1k}\quad v_{2k}\quad \cdots\quad v_{rk}]^{\mathrm{T}}$.

2)p 分位线 V^p 由所有分位点 v^{kp} 组成:

$$V^p=[v^{1p}\quad v^{2p}\quad \cdots\quad v^{kp}\quad \cdots\quad v^{rp}],v^{kp}=pct(V_k^*,p) \tag{5}$$

其中 $pct(V_k^*,p)$ 函数表示计算 V_k^* 向量的 p 分位数.

3)近邻特征曲线在 p 分位线上方的判定条件如下:

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(v_k^s>v_k^p)>f \tag{6}$$

其中 f 是覆盖比例,即某样本近邻特征曲线上的点超过 p 分位数特征曲线对应点的比例大于阈值 f 时,认为该样本属于真噪声集 D_{TN}.

使用分位线作为参考值主要是考虑到分位线对极端值不敏感,而且分位线能够自动根据数据之间的相对距离灵活调整噪声识别标准,不受数据尺度变化影响. 总体上,通过比较近邻特征曲线,可以进一步排除类簇分布不理想所产生的假噪声,从而提高噪声检测的准确性.

2.4　标签噪声检测与清洗算法

借助无监督聚类技术可以将大多数标签噪声检测出来,但难以保证所有可疑标签均为噪声. 为提升标签噪声检测的准确率,本文提出一种基于类簇分布的噪声检测算法,通过特征自适应增强方法提高类簇分布和观测类标签的语义一致性,同时利用近邻特征曲线排除可疑噪声中的虚假标签噪声,

最后对噪声数据集进行标签清洗以提升标签质量. 具体步骤如算法 1 所示.

算法 1. 标签噪声检测与清洗算法

输入: 带标签噪声的数据集 $D=\{x_i,y_i\}_{i=1}^{n}$,特征曲线分位数 p,覆盖比例 f.

输出: 清洗后数据集 D^*.

1. 对数据集 D 进行 5 折交叉验证,取多个分类模型在第 c 类上的平均准确率作为 Acc_c;
2. 利用式(3)计算各类的权重 w_c;
3. 按照式(2)利用权重 w_c 和标签矩阵 H 对属性矩阵 X 进行增强,并对增强后的属性集合 X′进行层次聚类;
4. 将每个类簇中的多数观测类标记作为该类簇的标签,对比每个样本的观测类标签与所在类簇的标签. 如果两个标签不一致,则将该样本列入可疑噪声集 D_N,否则将其列入干净集 D_C;
5. 利用式(4)计算出的行向量 V_s 作为对应样本的近邻特征曲线,并利用式(5)计算 p 分位曲线,将满足式(6)条件的样本列入真噪声集 D_{TN},不满足式(6)条件的样本列入假噪声集 D_{FN};
6. if $C=2$
7. 将 D_{FN} 中的样本标签反转纠正,得到清洗后数据集 D^*;
8. else
9. 对并集 $D_C\cup D_{TN}$ 重复执行 5 次交义验证;若某噪声样本在 5 次结果中有 4 次及以上的预测标签均一致,则将此样本对应标签纠正为多数标签,所有纠正标签后的样本构成纠正集 D_T,最后返回清洗后的 $D^*=D_C\cup D_T$.
10. end

算法 1 的大致流程如图 3 所示,主要包括特征增强、噪声检测和噪声清洗 3 个部分. 第 1 部分左侧区域对应数据增强的步骤 1 ~ 3;中间区域表示噪声检测过程,对应步骤 4 ~ 5;右侧区域将纠正集和干净集合并形成最终清洗后的数据集,对应步骤 6 ~ 10. 当类别数较多时,标签的精准纠正可能会变得较为困难,因此步骤 9 利用多数投票原则纠正比较确信的噪声标签样本,放弃比较模糊的噪声标签样本,以避免出现错误纠正的情况.

算法 1 的时间复杂度主要取决于聚类过程和模型预测过程,其中层次聚类有最高的时间复杂度为 $O(n^3)$,因此该算法的时间复杂度 $T(CDGDC)=O(n^3)$.

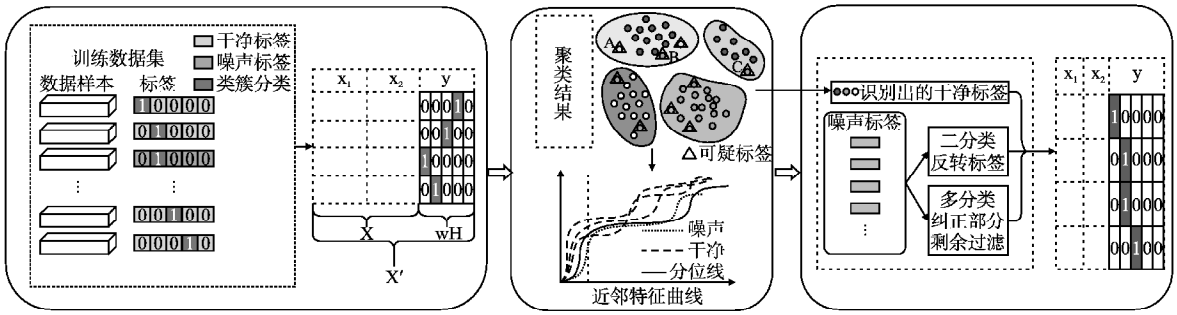


图 3 类簇分布引导的标签噪声检测框架

Fig. 3 Framework for label noise detection guided by cluster distribution

3 实验结果与分析

本节介绍噪声检测算法在标准数据集上的实验设置、噪声识别结果、泛化性能分析以及参数敏感性分析,然后在人脸性别标签数据集上验证所提算法的有效性.

3.1 实验设置

本文实验在 20 个二分类和 10 个多分类数据集上评估噪声检测算法的性能. 为检测各算法的噪声识别能力,在原干净数据集中随机选择 80% 的样本作为训练集,剩余的 20% 作为测试集. 给训练集中加入不同比例(NR = 0.1,0.15,0.2,0.3)的人工噪声以模仿真实含噪场景下的数据标签. 实验中选取 ANN、CF、IPF、MVF、CRF、DRNC 作为对比算法,同时对比了不作数据处理(Nothing)的基准情况. 将噪声检测出来之后执行反转(二分类数据集)或过滤(多分类数据集)处理,最后将分类器在清洗后的数据集上训练,并在测试集上评价分类器的泛化性能. 为了降低实验的随机性,所有实验重复 5 次取平均值作为最后的结果.

表 1 列出了实验所用的数据集信息,包括其样本量、特征数和类别数.

对比算法的参数根据已有文献设置如下:ANN 中的近邻数 K 值设为 3;CF 采用 C4.5 为基分类器,进行 10 折交叉验证;IPF 采用 C4.5 为基分类器,且其中划分的子集数设置为 5;MVF 采用最近邻、C4.5 和朴素贝叶斯作为 3 个基分类器,且其中划分的子集数设置为 4;CRF 中随机树数量设为 100,过滤阈值设为 5,上述参数均为原文中的推荐参数设置,本文所提算法的聚类簇数由样本量和总类数确定,即 $a=\lceil\sqrt{n\times C}\rceil$,特征曲线的近邻数 K 取 20,在特征增强矩阵的权值设定中,准确率来自 3 个分类器的预测结果均值,包括逻辑回归分类器(LR)、随机森林(RF)和支持向量机(SVM). 选取近邻特征曲线的分位数 p 为 30% 和覆盖比例 f 为 70% 进行实验.

在噪声检测阶段,使用噪声检测准确率(DAcc)评价算法效果:

$$DAcc=\frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

其中,TP 表示被正确识别为噪声的误标记样本数,FP 表示被错误识别为噪声的干净样本数,TN 表示被正确识别为无噪的干净样本数,FN 表示被错误识别为无噪的误标记样本数.

在泛化性能测试阶段,实验采用 K 近邻(KNN)、逻辑回归(LR)、决策树(DT)、神经网络(NNet)、支持向量机

(SVM)、随机森林(RF)分类器的测试准确率来评价泛化性能.

表 1　数据集信息

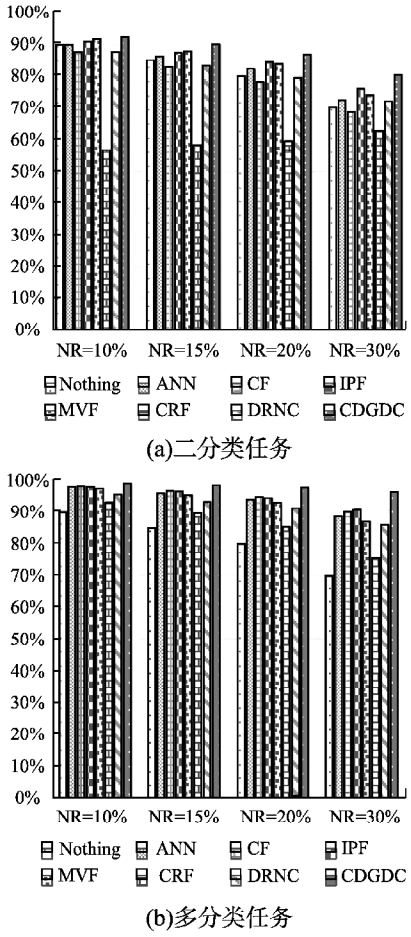
Table 1　Datasets information

数据集编号	数据集	样本量	特征数	类别数
1	Sonar	208	60	2
2	Ionosphere	351	34	2
3	Votes	435	16	2
4	Musk	476	166	2
5	Climate	540	18	2
6	Australian	690	14	2
7	Breast_cancer	699	10	2
8	diabetes	768	8	2
9	Fourclass	862	2	2
10	GermanNumer	1000	24	2
11	Diabetic	1151	20	2
12	Svmguide3	1243	22	2
13	Banknote	1372	5	2
14	Titanic	2201	3	2
15	Splice	3175	60	2
16	Svmguide1	7089	4	2
17	Mushrooms	8124	112	2
18	phishing	11055	68	2
19	EEG Eye State	14980	15	2
20	MAGICGammaTelescope	19020	11	2
21	Seeds	210	7	3
22	Glass Identification	214	10	6
23	Svmguide2	391	20	3
24	Svmguide4	612	10	6
25	Vehicle	846	18	4
26	Phishing	1353	10	3
27	Cardiotocography	2126	23	3
28	Segment	2310	19	7
29	Abalone	4177	8	29
30	Letter	20000	16	26

3.2　噪声识别

考虑到多分类任务的复杂性,所提 CDGDC 算法对多类标签噪声的清洗部分与二分类的处理有所区别,因此实验部分对二分类和多分类任务分别进行对比分析.

图 4 画出了二分类和多分类任务下的噪声检测准确率.整体上看在多分类任务下的噪声检测准确率普遍高于二分类任务,可能是因为类别数较多时同类样本中会出现多个类别的噪声,这样样本间差异更明显也就相对更容易识别.从噪声比例的角度来看,8 种方法都是噪声比例越高,识别准确率就越低,本文所提 CDGDC 算法的识别准确率随噪声比例的下降速度相对更缓慢;在各种噪声比例下,CDGDC 算法的噪声检测准确率最高,相较于基准方法(Nothing)的提升幅度最大.图 4(a)中 CRF 算法的噪声检测准确率低于基准方法,可能是因为阈值缺乏自适应性影响了噪声识别的目标函数,进而导致出现过度清洗.图 4(b)中基准方法(Nothing)准确率最低,CRF 和 DRNC 的检测准确率略高于基准方法,ANN、MVF、CF 和 IPF 的噪声检测准确率比较接近且仅次于 CDGDC.在多分类任务中,样本分布在更高维的类别空间中,噪声样本不仅与一个类别的分布不一致,而且可能与多个类别的样本分布都不匹配,因此会提高分类器对噪声的检测能力.图 4 中本文 CDGDC 算法通过真假噪声识别方法进一步提升了噪声检测准确率,在二分类和多分类任务下均为最优.



(a)二分类任务

(b)多分类任务

图 4　不同噪声比例下的检测准确率

Fig. 4　Detection accuracy at different noise ratios

本文还使用 ROC 空间比较不同噪声清洗算法,ROC 空间的真正例率 TPR = TP/(TP + FN),假正例率 FPR = FP/(FP + TN).图 5 给出了不同噪声比例下各个清洗算法在所有数据集上的平均 TPR 和 FPR 值.图中参考线是关于 TPR 和 FPR 的调和平均函数$\frac{2\times TPR(1-FPR)}{TPR+1-FPR}$.理想的过滤算法应该靠近 ROC 空间的左上角,即调和平均函数值越大越好.

从图 5 中可以看出,不同噪声比例下 DRNC 算法均在 ROC 空间下方,表明 DRNC 算法将噪声样本错误识别为干净的样本数量(FN)较多,即很多噪声漏检使得 TPR 值偏低;CRF 算法将很多的无噪样本错误识别为噪声(FP 较大),即很多噪声误检使得 FPR 值偏高.

漏检和误检都会导致检测准确率偏低,即算法识别噪声的效果差.由图 5 调和平均参考线位置可知,当噪声比例为 0.1、0.15 和 0.2 时,IPF、CF 和 ANN 在本文所提方法中识别效果较好(对应参考线在 0.7 附近),当噪声比例为 0.3 时,

IPF、CF、CRF 相对较好(对应参考线在 0.6 以上). CDGDC 算法的 TPR 值最高且 FPR 值最低,位于调和平均值为 0.8 的参考线的左上方,表明 CDGDC 的调和平均结果最高,其综合噪声检测性能最好.

3.3 泛化性能分析

表 2 和表 3 分别在二分类和多分类数据上对比了 8 种噪声清洗方法对泛化性能的影响,其中每行最大值加粗,排名前三的数值加下划线由表 2 可以看出,噪声比例越高,分类准确

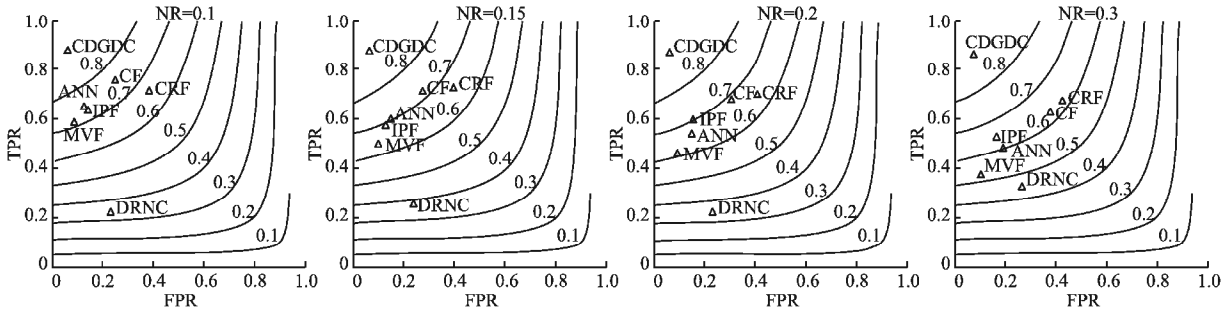


图 5 ROC 空间中的噪声检测

Fig. 5 Noise detection in the ROC space

表 2 二分类数据清洗后的各分类器预测准确率

Table 2 Predictive accuracy of each classifier after binary data cleaning

Noise	model	Nothing	ANN	CF	IPF	MVF	CRF	DRNC	CDGDC
NR = 10%	SVM	0.805	0.803	0.807	0.806	**0.808**	0.596	0.797	0.806
	KNN	0.801	0.798	0.793	0.808	0.809	0.607	0.798	**0.812**
	DT	0.775	0.793	0.768	0.796	0.799	0.606	0.771	**0.806**
	RF	0.834	0.832	0.827	0.836	0.841	0.589	0.827	**0.842**
	NNet	0.796	0.810	0.794	0.810	0.808	0.602	0.797	**0.825**
NR = 15%	SVM	0.803	0.801	0.801	**0.805**	0.804	0.605	0.792	**0.805**
	KNN	0.784	0.786	0.775	0.787	0.797	0.609	0.775	**0.807**
	DT	0.745	0.772	0.737	0.781	0.777	0.618	0.752	**0.794**
	RF	0.809	0.820	0.802	0.821	0.825	0.602	0.798	**0.839**
	NNet	0.780	0.794	0.773	0.793	0.791	0.609	0.773	**0.811**
NR = 20%	SVM	0.793	0.793	0.796	0.793	**0.797**	0.616	0.790	0.794
	KNN	0.755	0.761	0.748	0.778	0.782	0.619	0.755	**0.792**
	DT	0.710	0.738	0.702	0.764	0.754	0.629	0.717	**0.782**
	RF	0.786	0.797	0.777	0.810	0.811	0.615	0.780	**0.823**
	NNet	0.753	0.773	0.756	0.779	0.772	0.626	0.757	**0.800**
NR = 30%	SVM	0.772	0.768	0.773	0.778	**0.780**	0.628	0.763	0.774
	KNN	0.663	0.678	0.658	0.722	0.696	0.645	0.705	**0.741**
	DT	0.640	0.676	0.636	0.716	0.684	0.648	0.670	**0.750**
	RF	0.721	0.737	0.709	0.755	0.751	0.637	0.739	**0.791**
	NNet	0.699	0.726	0.693	0.728	0.722	0.651	0.713	**0.764**
平均值		0.761	0.773	0.756	0.783	0.780	0.618	0.763	**0.798**
最佳次数		0	0	0	2	3	0	0	**17**
前三次数		0	4	2	19	16	0	1	**20**

率总体上越低;在相同噪声比例和相同的预测分类器下,所提 CDGDC 算法在大多数情况下(17/20 = 85%)得到最高的分类准确率;从排名前三出现的次数来看,CDGDC 在所有情况下均在前三,次高的是 IPF 算法(19/20),但 CDGDC 平均比 IPF 高 1.5%,说明本文所提清洗方法对泛化性能提升效果明显;从预测分类器的角度来看,SVM 和 RF 分类器的准确率最高,说明两者对噪声数据相对更为鲁棒.

由表 3 可以看出,噪声比例越高,分类准确率总体上越低;在相同噪声比例和相同的预测分类器下,所提 CDGDC 算法在大多数情况下(16/20 = 80%)得到最高的分类准确率,且 CDGDC 的准确率比次高的 MVF(16/20)平均高 1%,说明本文所提清洗方法对泛化性能提升更有效. 综上可见,在各种分类任务和不同噪声比例下,本文的清洗算法能够最大程度地提升多种常见分类模型的泛化能力.

图 6 给出了各分类模型预测准确率的临界差异图,其中算法排名越靠前表示预测准确率越高. 该图显示了算法差异的 Nemenyi 统计检验结果,如果各算法之间的排名距离不超过临界差异值 CD,表明算法之间的差异不显著,通过横线连接. 算法的平均排名基于 30 个数据集和 4 种噪声比例的预测准确率计算得出. 由图 6 可知,除 CRF 外的清洗算法都能提

升分类器的泛化性能,本文 CDGDC 算法提升效果显著优于其他算法. 从预测分类器来看,SVM 和 RF 分类器在算法差异上的结果相似,而其他分类器如 IPF、MVF、ANN 之间差异并不显著,且其泛化性能处于中等偏上水平. 本文 CDGDC 算法在 6 种分类模型上都取得最优预测准确率排名,泛化性能最好,主要得益于本文算法的噪声识别准确率最高.

表 3　多分类数据清洗后的各分类器预测准确率

Table 3　Predictive accuracy of each classifier after multi-class data cleaning

Noise	model	Nothing	ANN	CF	IPF	MVF	CRF	DRNC	CDGDC
NR = 10%	SVM	0.690	0.693	0.696	0.692	0.697	0.635	0.702	**0.709**
	KNN	0.703	0.711	0.713	0.708	0.717	0.654	0.701	**0.731**
	DT	0.685	0.723	0.705	0.691	**0.725**	0.650	0.684	0.706
	RF	0.710	0.672	0.648	0.587	0.726	0.579	0.711	**0.730**
	NNet	0.709	0.729	0.735	0.722	**0.731**	0.677	0.713	0.730
NR = 15%	SVM	0.694	0.683	0.697	0.692	0.684	0.615	0.698	**0.706**
	KNN	0.675	0.692	0.711	0.710	0.708	0.628	0.676	**0.725**
	DT	0.664	0.699	0.709	0.692	**0.711**	0.619	0.681	0.707
	RF	0.700	0.651	0.654	0.640	0.701	0.590	0.702	**0.714**
	NNet	0.688	0.709	0.723	0.716	0.719	0.642	0.684	**0.725**
NR = 20%	SVM	0.690	0.689	0.682	0.679	0.691	0.613	0.696	**0.701**
	KNN	0.642	0.678	0.697	0.700	0.695	0.618	0.646	**0.716**
	DT	0.652	0.681	0.688	0.681	**0.706**	0.619	0.642	0.698
	RF	0.686	0.653	0.646	0.599	0.687	0.582	0.702	**0.707**
	NNet	0.655	0.690	**0.713**	0.702	0.709	0.638	0.674	**0.713**
NR = 30%	SVM	0.641	0.664	0.672	0.642	0.669	0.623	0.668	**0.680**
	KNN	0.577	0.637	0.671	0.672	0.652	0.604	0.591	**0.685**
	DT	0.569	0.643	0.650	0.653	0.661	0.596	0.584	**0.671**
	RF	0.654	0.631	0.630	0.566	0.669	0.566	0.656	**0.686**
	NNet	0.616	0.645	0.658	0.673	0.670	0.635	0.631	**0.679**
平均值		0.665	0.679	0.685	0.671	0.696	0.619	0.672	**0.706**
最佳次数		0	0	1	0	4	0	0	**16**
前三次数		0	2	10	5	16	0	7	**20**

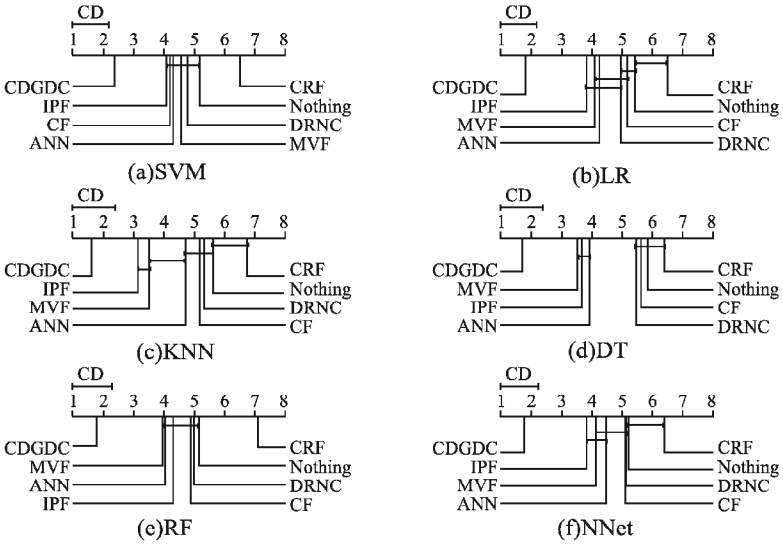


图 6　各分类模型预测准确率的临界差异图

Fig. 6　CD diagram of prediction accuracy for each classification model

3.4　参数敏感性分析

在标签噪声检测过程中,当特征分位数 p 和覆盖比例 f 选择不同时,CDGDC 算法的噪声检测结果有所差异. 为取得更好的识别效果,本文对特征分位数和覆盖比例进行参数敏感性分析. 在标准数据集中人工加噪以模拟真实数据集,选择特征分位数分别为 10%、20%、30%、40%、50%,覆盖比例分别为 60%、65%、70%、75%、80% 进行实验. 为获得相对稳定可靠的结果,每次实验重复 5 次并取其均值作为最后结果.

图 7 列出了不同的分位数和覆盖比例的组合下 CDGDC

算法的噪声检测准确率,横轴和纵轴分别表示分位数和覆盖比例,图中数值表示噪声检测准确率.从图7可以看出,在分

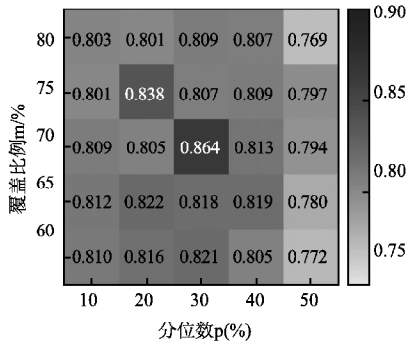


图 7 参数敏感性分析

Fig. 7 Parameter sensitivity analysis

位数 p 为 30%、覆盖比例 f 为 70% 的情况下噪声检测准确率相对最高,能够最有效地筛选出真实噪声,故本文方法采用此参数设置.

3.5 真实性别数据集噪声检测

实验使用真实数据集 wiki[25] 进行测试,人脸图片共 62328 张,且其都有对应的性别标签,但由于多人标注的主观性导致性别标签与图像不一定匹配,即存在大量标签噪声.此外数据集中有 2463 个图像未标注,本文在该数据集上进行标签噪声识别与清洗,拟找出性别标签与图片不符合的情况,并对比在不同分类器下的分类准确率.

利用 VGG-16 深度模型从图像中提取特征[26],同时给所有空标签样本赋予随机标记.将 wiki 均匀划分为 5 个子集,每次选择其中一个子集作为验证集,其余 4 个子集作为训练集进行清洗并训练分类器,重复上述过程 5 次并取其平均值作为最终结果.表 4 列出了不同清洗算法在 4 种分类模型下的分类准确率,其中每行最大值加粗,排名前三的数值加下划线.由排名前三的次数可以看出 MVF 和 CRF 的分类准确率在已有方法中较高.相对其他方法,CDGDC 能够提升分类器的准确率,且该方法的测试准确率最高.

表 4 真实性别数据集在不同分类器下的分类准确率

Table 4 Classification accuracy of real gender data sets with different classifiers

模型	Nothing	ANN	CF	IPF	MVF	CRF	DRNC	CDGDC
LR	0.691	0.688	0.669	0.696	0.748	0.743	0.719	**0.770**
KNN	0.670	0.668	0.656	0.676	0.735	0.738	0.708	**0.763**
DT	0.616	0.639	0.605	0.630	0.698	0.711	0.657	**0.757**
RF	0.751	0.739	0.734	0.754	0.764	0.760	0.757	**0.780**
前三次数	0	0	0	0	4	4	0	**4**

图 8 显示了 CDGDC 算法所检测出的部分标签噪声,各图片下方列出了对应的图像名称、原始性别标签以及 CDGDC 的纠正标签.由图可见,本文所提方法对已有标记数据和随机标记数据都能够准确识别并纠正性别标签噪声,实验结果验证了其有效性.所提方法可以提高真实数据的标签质量,增强分类器的泛化性能.

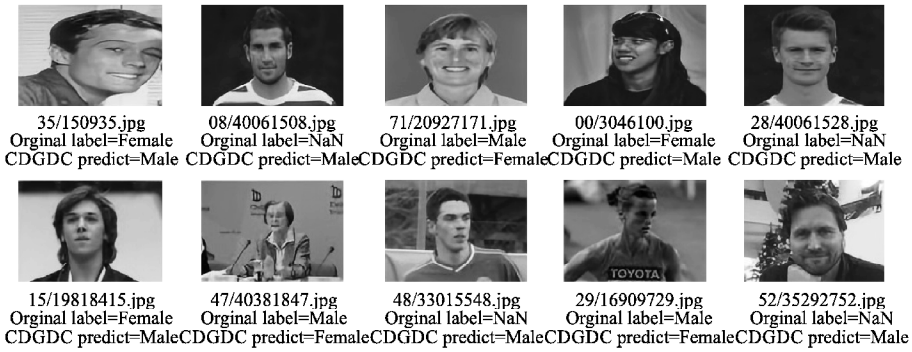


图 8 性别标签噪声的 CDGDC 清洗结果

Fig. 8 CDGDC cleaning results of noisy gender labels

4 结束语

已有噪声清洗方法不仅能够识别出数据集中的标签噪声,还能够在一定程度上辅助纠正这些噪声.然而这些方法没有充分考虑数据分布信息,导致噪声识别准确性不够高.本文提出一种基于类簇分布的标签噪声检测方法,它通过类簇中的少数类估计潜在标签噪声,然后利用类别属性增强方法提升类标签和类簇之间的语义一致性,同时使用近邻特征曲线进一步排除伪标签噪声.实验结果表明该算法具有更好的灵活性和适应性,能够在不同噪声水平下更准确地识别标签噪声,对提升分类器泛化性能和预测可靠性具有重要意义.

本文在提出的识别标签噪声方法的基础上,对不同分类任务采用过滤和纠正结合的清洗方案对噪声样本进行处理,实验结果证明该方案的效果良好,同时通过排除伪噪声能防止过度清洗.下一步,可以将本文的标签噪声检测与清洗算法拓展到深度学习模型以及回归任务中,进一步修正噪声检测的方法和清洗策略,以提升数据的质量和分类器的泛化能力.

References:

[1] Song H, Kim M, Park D, et al. Learning from noisy labels with deep neural networks: a survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(11): 8135-8153.

[2] Jiang G X, Wang W J, Qian Y H, et al. A unified sample selection framework for output noise filtering: an error-bound perspective [J]. Journal of Machine Learning Research, 2021, 22(18): 1-66.

[3] Kovashka A, Russakovsky O, LI F F, et al. Crowdsourcing in computer vision[J]. Foundations and Trends in Computer Graphics and Vision, 2016, 10(3): 177-243.

[4] Chen Q Q, Jiang G X, Cao F Y, et al. A general elevating framework for label noise filters[J]. Pattern Recognition, 2024, 147: 110072, doi: 10.1016/j.patcog.2023.110072.

[5] MEN C Q, MENG X C, JIANG G X, et al. Active label noise cleaning method based on spxy sampling[J]. Journal of Chinese Computer Systems, 2021, 42(9): 1865-1870.

[6] Lienen J, Hullermeier E. Mitigating label noise through data ambiguation[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 13799-13807.

[7] Jiang G X, Zhang J, Bai X, et al. Which is more effective in label noise cleaning, correction or filtering? [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 12866-12873.

[8] Shu J, Xie Q, Yi L X, et al. Meta-weight-net: learning an explicit mapping for sample weighting[C]//Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press, 2019: 1917-1928.

[9] JIANG G X, WANG W J. A numerical label noise filtering algorithm for regression[J]. Journal of Computer Research and Development, 2022, 59(8): 1639-1652.

[10] Jiang L, Zhou Z, Leung T, et al. MentorNet: learning data-driven curriculum for very deep neural networks on corrupted labels [C]//International Conference on Machine Learning, Proceedings of Machine Learning Research(PMLR), 2018: 2304-2313.

[11] Zhang W N, Tan X Y. Combining outlier detection and reconstruction error minimization for label noise reduction[C]//IEEE International Conference on Big Data and Smart Computing (BigComp), 2019: 1-4.

[12] Huang J, Zhu Q, Yang L, et al. A non-parameter outlier detection algorithm based on Natural Neighbor[J]. Knowledge-Based Systems, 2016, 92(15): 71-77.

[13] WANG M, WU W J, LIU H Y, et al. Confidence prediction methods for noisy label identification and correction[J]. Journal of Northwestern University, 2022, 52(5): 857-867.

[14] Xia S, Wang G, Chen Z, et al. Complete random forest based class noise filtering learning for improving the generalizability of classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(11): 2063-2078.

[15] Cao J, Kwong S, Wang R. A noise-detection based AdaBoost algorithm for mislabeled data[J]. Pattern Recognition, 2012, 45(12): 4451-4465.

[16] Saez J A, Galar M, Luengo J, et al. INFFC: an iterative class noise filter based on the fusion of classifiers with noise sensitivity control [J]. Information Fusion, 2016, 27: 19-32, doi: 10.1016/j.inffus.2015.04.002.

[17] HOU S Y, JIANG G X, WANG W J. Label noise filtering method based on relative outliers[J]. Journal of Automation, 2024, 50(1): 154-168.

[18] FAN R X, JIANG G X, WANG W J. An outlier detection algorithm for personalized k-nearest neighbors[J]. Journal of Chinese Computer Systems, 2020, 41(4): 752-757.

[19] Yi K, Wu J X. Probabilistic end-to-end noise correction for learning with noisy labels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2019: 7017-7025.

[20] Bhandari K, Kumar K, Sangal A L. Data quality issues in software fault prediction: a systematic literature review[J]. Artificial Intelligence Review, 2023, 56(8): 7839-7908.

[21] XU M L, JIANG G X, WANG W J. A label noise filtering framework based on anomaly detection[J]. Computer Science, 2024, 51(2): 87-99.

[22] Lee D, Kim K. Improved noise-filtering algorithm for AdaBoost using the inter-and intra-class variability of imbalanced datasets[J]. Journal of Intelligent & Fuzzy Systems, 2022, 43(4): 5035-5051.

[23] Zhang J, Jiang X, Tian N, et al. Label noise correction for crowdsourcing using dynamic resampling[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108439, doi: 10.1016/j.engappai.2024.108439.

[24] Yuan W W, Guan D H, Ma T H, et al. Classification with class noises through probabilistic sampling[J]. Information Fusion, 2018, 41: 57-67, doi: 10.1016/j.inffus.2017.08.007.

[25] Rothe R, Timofte R, Van G L. Deep expectation of real and apparent age from a single image without facial landmarks[J]. International Journal of Computer Vision, 2018, 126(2): 144-157.

[26] Muhammad U, WANG W, Chattha S P, et al. Pre-trained VGGNet architecture for remote-sensing image scene classification[C]// Proceedings of the 24th International Conference on Pattern Recognition(ICPR), 2018: 1622-1627.

附中文参考文献:

[5] 门吕骞,孟晓超,姜高霞,等. 一种利用 SPXY 采样的标签噪声主动清洗方法[J]. 小型微型计算机系统, 2021, 42(9): 1865-1870.

[9] 姜高霞,王文剑. 面向回归任务的数值型标签噪声过滤算法[J]. 计算机研究与发展, 2022, 59(8): 1639-1652.

[13] 汪　敏,伍文静,刘瀚阳,等. 噪声标签识别与纠正的置信度预测方法[J]. 西北大学学报, 2022, 52(5): 857-867.

[17] 侯森寓,姜高霞,王文剑. 基于相对离群因子的标签噪声过滤方法[J]. 自动化学报, 2024, 50(1): 154-168.

[18] 樊瑞宣,姜高霞,王文剑. 一种个性化 k 近邻的离群点检测算法[J]. 小型微型计算机系统, 2020, 41(4): 752-757.

[21] 许茂龙,姜高霞,王文剑. 基于异常检测的标签噪声过滤框架[J]. 计算机科学, 2024, 51(2): 87-99.