基于相对离群因子的标签噪声过滤方法

侯森寓 [1] 姜高霞 [1] 王文剑 [1,2]

摘 要 分类任务中含有类别型标签噪声是传统数据挖掘中的常见问题, 目前还缺少针对性方法来专门检测类别型标签噪声. 离群点检测技术能用于噪声的识别与过滤, 但由于离群点与类别型标签噪声并不具有一致性, 使得离群点检测算法无法精确检测分类数据集中的标签噪声. 针对这些问题, 提出一种基于离群点检测技术、适用于过滤类别型标签噪声的方法 —— 基于相对离群因子 (Relative outlier factor, ROF) 的集成过滤方法 (Label noise ensemble filtering method based on relative outlier factor, EROF). 首先, 通过相对离群因子对样本进行噪声概率估计; 然后, 再迭代联合多种离群点检测算法, 实现集成过滤. 实验结果表明, 该方法在大多数含有标签噪声的数据集上, 都能保持优秀的噪声识别能力, 并显著提升各种分类模型的泛化能力.

关键词 分类, 标签噪声, 离群点检测, 相对离群因子, 噪声过滤

引用格式 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. 自动化学报, 2024, **50**(1): 154−168

DOI 10.16383/j.aas.c230117

A Label Noise Filtering Method Based on Relative Outlier Factor

HOU Sen-Yu[1] JIANG Gao-Xia[1] WANG Wen-Jian[1,2]

Abstract The presence of categorical label noise in classification tasks is a common issue in traditional data mining. Currently, there is a lack of targeted methods specifically designed to detect categorical label noise. While outlier detection techniques can be used for noise identification and filtering, the lack of consistency between outliers and categorical label noise renders outlier detection algorithms unable to accurately detect label noise in classification data sets. To address these issues, a method based on outlier detection techniques, called the label noise ensemble filtering method based on relative outlier factor (ROF) (EROF), is proposed for filtering categorical label noise. The EROF method estimates noise probability of samples by using relative outlier factor and then iteratively combinings multiple outlier detection algorithms for ensemble filtering. Experimental results show that this method maintains excellent noise identification capability in most data sets which contain label noise, and significantly improves the generalization ability of various classification models.

Key words Classification, label noise, outlier detection, relative outlier factor (ROF), noise filtering

Citation Hou Sen-Yu, Jiang Gao-Xia, Wang Wen-Jian. A label noise filtering method based on relative outlier factor. *Acta Automatica Sinica*, 2024, **50**(1): 154−168

分类是机器学习领域中一项重要的任务, 大量研究表明, 数据质量决定着训练出的分类模型的泛化性能[1]. 随着人工智能技术的进步, 风险与挑战也随之而来, 许多人工智能应用领域 (如医学诊断、人脸识别和智能驾驶等) 需要更高的数据质量, 以保证模型的准确率 (Accuracy, Acc)[2−4]. 然而, 由于对数据实施可靠标记通常是昂贵而耗时的, 对于实际的机器学习来说, 数据普遍包含噪声是一个不容忽视的问题[5]. 因此, 实施噪声清除或降低噪声对模型泛化性能的影响是十分必要的.

监督学习中的数据噪声主要分为特征噪声和标签噪声两种类型[6]. 特征噪声是指观测到的特征与真实特征存在误差的数据, 例如由于物联网设备故障, 部分数据采集到错误、缺失或不完整的特征值[7]; 标签噪声是指观测到的实例标签与真实标签不一致的数据, 例如在医疗诊断中, 由于专家标记的不一致性, 造成病例数据标签错误[8]. 文献 [9] 从统计角度考量, 提出标签噪声主要有完全随机噪声、随机噪声和非随机噪声三类. 它们通过噪声样本与其特征和标签的关联性进行区分, 完全随机噪声的产生

收稿日期 2023-03-08 录用日期 2023-06-14
Manuscript received March 8, 2023; accepted June 14, 2023
国家自然科学基金 (U21A20513, 62276161, 62076154), 中央引导地方科技发展基金 (YDZX20201400001224), 山西省 1331 工程重点学科建设基金资助
Supported by National Natural Science Foundation of China (U21A20513, 62276161, 62076154), Central Government Guides Local Science and Technology Innovation Projects (YDZX20201400001224), and Fund for the Shanxi Province 1331 Project of Key Subjects Construction
本文责任编委 张敏灵
Recommended by Associate Editor ZHANG Min-Ling
1. 山西大学计算机与信息技术学院 太原 030006 2. 计算智能与中文信息处理教育部重点实验室 太原 030006
1. School of Computer and Information Technology, Shanxi University, Taiyuan 030006 2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006

独立于数据样本的特征值和标签类别; 随机噪声和非随机噪声的产生与数据的特征或标签类别之间存在一定的关联性, 这种关联可能是由于标签分配过程中的系统性、主观性错误或模糊类别边界等原因造成的.

根据机器学习任务的不同, 标签噪声在分类与回归任务中, 又分别称为类别型标签噪声和数值型标签噪声[9]. 在分类任务中, 无论是特征噪声还是类别型标签噪声都会影响模型的泛化性能, 但研究表明, 标签噪声比特征噪声具有更大的危害性. 一方面, 数据通常具有多个特征, 而标签是唯一的; 另一方面, 每个特征对于分类模型学习的重要性不尽相同, 而标签的正确与否对模型学习有着更大的影响. 因此, 处理分类任务中的标签噪声问题显得至关重要.

解决分类任务中的标签噪声问题可以从算法层和数据层 2 个层面考虑. 在算法层面的处理依赖于构建对标签噪声具有鲁棒性的模型 (如重构损失函数和加权集成等方式), 通过提高模型对噪声的容忍, 以减少其受标签噪声的影响. 但是, 鲁棒模型并不能完全鲁棒[10]. 数据层面的处理是通过标签噪声过滤的方法提高数据质量[11], 标签噪声在训练前就已经被识别并处理掉. 显然, 标签噪声过滤更直接有效.

针对分类任务中的类别型标签噪声问题, 本文从离群点与噪声的相关性出发, 提出相对离群因子 (Relative outlier factor, ROF), 以评估数据的噪声概率, 并据此构建标签噪声过滤方法. 本文主要贡献有以下 3 个方面:

1) 基于离群点检测算法提出相对离群因子, 它可以对数据每个样本点进行噪声概率评估, 依据该评估, 可以实现噪声检测和初步的过滤;

2) 以提高模型泛化能力为目标, 根据 1) 提出的噪声概率评估方法, 设计基于相对离群因子的集成过滤方法 (Label noise ensemble filtering method based on relative outlier factor, EROF), 使得噪声过滤更加稳定精确;

3) 在标准数据集的实验结果表明, 本文方法与现有的噪声过滤方法相比, 能更有效地提升数据质量和模型的泛化性能.

1 相关工作

本节主要介绍已有的标签噪声过滤方法和离群点检测算法, 并对这些方法进行简单归纳和总结.

1.1 标签噪声过滤方法

标签噪声过滤方法一般是指检测并过滤训练数据中的噪声样本, 通过该方法提升数据质量, 以确保训练出的模型性能.

早期过滤算法是利用分类模型对噪声的敏感性来检测标签噪声. 常见做法是将分类器对样本的预测标签和样本观测标签的一致性作为识别标签噪声的指标之一[12]. 这个方法被称为分类过滤, 文献 [13] 提出运用 K 折交叉验证方法, 将测试集上错误分类的样本直接视为噪声删除. 然而, 分类过滤面临着“先有鸡, 还是先有蛋”的悖论[14], 该过滤方法必须有精确的分类器, 而在含噪数据集上, 大概率训练出较差的分类器.

由于集成学习的广泛应用, 基于不同集成策略的过滤方法应运而生, 代表性方法有多数投票过滤器 (Majority vote filter, MVF)[15]、动态集成过滤器[16] 和高一致性随机森林 (Random forest, RF) 过滤器[17]. 这些过滤方法的主要思想都是通过多个基分类器的预测结果组合后的正确程度, 来识别噪声. 集成过滤器通常比单一基过滤器具有更好的精度, 但其本质上还是基于模型预测的过滤方法, 仍存在上述的悖论问题, 且计算成本通常更大.

更为常见的是基于近邻模型的过滤方法, 通常需要借助 K 近邻 (K-nearest neighbors, KNN) 模型实现. 如编辑近邻[18] 过滤器、全近邻[19] 过滤器、互近邻 (Mutual nearest neighbor, MNN)[20] 过滤器. 这些过滤器对近邻参数 k 的选取过于敏感. 近邻感知[21] 过滤算法采用迭代搜索思想, 解决参数 k 的自适应问题, 但仍存在一定的维度灾难问题.

在准确性和可靠性方面, 依赖分类器和近邻模型识别标签噪声的效果均不理想, Xia 等[22−24] 提出基于相对密度 (Relative density-based, RD) 的过滤方法和基于完全随机森林 (Complete random forest, CRF) 的过滤方法, 并扩展应用于多分类数据集的噪声过滤中. RD 利用样本的相对密度来衡量样本的噪声强度, CRF 通过构建完全随机树来衡量样本被异类样本包围的水平, 进而确定样本的噪声强度. 为解决其中的硬阈值问题, 基于自适应投票策略的相对密度过滤器 vRD[24] 算法和自适应完全随机森林过滤器 Adp_mCRF[25] 算法被相继提出. 但是, 这两种方法都使用随机划分测试集的分类精度作为自适应指标, 导致在噪声比例 (Noise ratio, NR) 较高时, 过滤效果下降.

基于深度学习的标签噪声过滤方法利用神经网络强大的表示学习能力捕获数据底层结构, 从而识别和过滤噪声. 例如, Lu 等[26] 依据神经网络对样本预测置信度进行噪声识别和过滤; Han 等[27] 基于深度神经网络对于噪声数据的记忆时效性, 提出“Co-teaching”深度学习范式, 此方法通过训练两个深度

神经网络对标签噪声数据进行遗忘, 缓解误差累积, 从而在高标签噪声环境下, 训练出更具鲁棒性的模型. 但基于深度学习的方法通常需要大量的计算资源和数据来训练模型, 数据量较少会导致模型过拟合, 无法泛化到新的数据.

Xiong 等[28] 基于离群点与噪声具有一定相似性的特点, 提出利用离群点检测技术, 完成噪声过滤的方法, 但该方法将离群点视为噪声的思想有失偏颇. Zhang 等[29] 提出基于稳健深度自动编码器的离群点检测技术过滤噪声, 该方法将检测出的离群点作为候选误标记数据, 再经过一种基于重构误差最小化方法, 验证最终视为噪声的样本, 但其本质仍是将样本是否为离群点作为噪声检测的基础.

1.2 离群点检测算法

离群点检测技术相较于类别型标签噪声过滤技术更为成熟, 出于简洁性、有效性、多样性等多方面考量, 本节简要介绍 6 种主流的离群点检测算法.

1) K 近邻[30] 离群点检测算法通过评估数据点与其第 k 近邻数据的距离, 来检测离群程度.

2) 局部离群因子 (Local outlier factor, LOF)[31] 检测算法利用局部邻域密度的概念, 来检测离群点.

3) 基于连通性离群因子[32] 检测算法在 LOF 算法基础上, 根据数据点的链接偏移程度, 来评价其离群度.

4) 为缓解维度灾难问题, 基于角度的离群点检测[33] 算法提出运用数据点与其近邻之间角度的加权方差, 来检测离群程度. 该方法能更精确地给出高维空间中数据点的离群因子.

5) 单类支持向量机 (Support vector machine, SVM)[34] 通过数据点到超平面边界的距离, 来计算离群值.

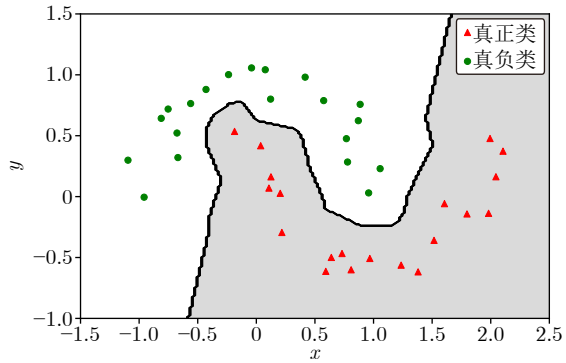
6) 基于近邻隔离的离群点算法[35] 使用孤立球模型, 实现离群因子的度量, 其中离群因子更高的数据点被半径更大的超球体所孤立.

2 基于相对离群因子的集成过滤方法

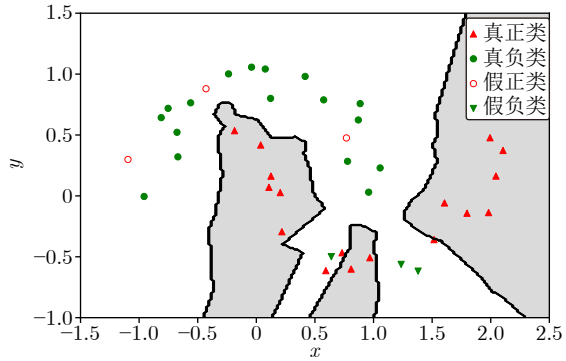
本节主要介绍相对离群因子的计算及应用, 并据此提出基于相对离群因子的集成过滤方法.

2.1 相对离群因子

为更形象地展示相对离群因子的相关概念与计算过程, 在人工生成的二分类数据集上进行模拟实验, 其中包括 34 个真实数据点和 6 个类别型标签噪声点, 选用 K 近邻分类 (K-nearest neighbors clasifier, KNNC) 模型作为基分类器 (近邻参数 k 设置为 1). 图 1 两个子图分别为在不含噪声数据集和包含噪声数据集中, 训练所得的分类模型决策边界图, 显然图 1(b) 的决策边界相对于图 1(a) 更复杂混乱, 并且图 1(b) 中部分处于决策边界的正常样本也被错分了, 这表明类别型标签噪声对分类模型产生一定负面影响. 因此, 有效处理类别型标签噪声是提高分类模型泛化能力的关键.



(a) 训练数据不含噪声
(a) Training data does not contain noise



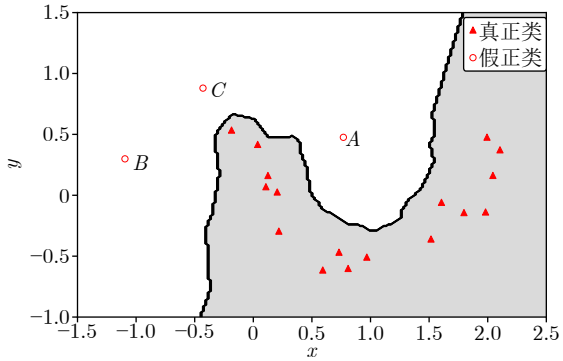
(b) 训练数据包含噪声
(b) Training data contains noise

图 1 不同含噪情况下的分类模型决策边界
Fig. 1 Decision boundary of classification model in different cases with noise

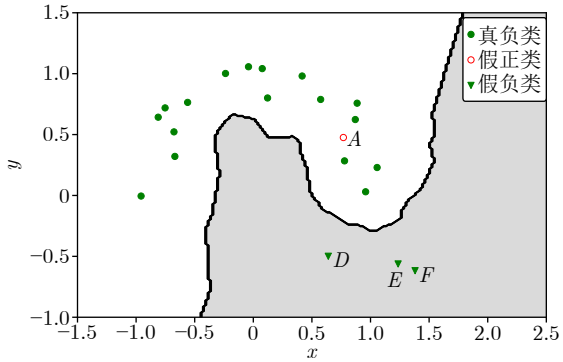
定义 1. 同质样本与异质样本

设有二分类数据集 $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, 其中 $\boldsymbol{x}_i \in \mathbf{R}^m$, $y_i \in \{$正类, 负类$\}$. 定义任意样本点的同质样本为观测标签与其相同的样本, 而异质样本为观测标签与其不同的样本.

图 2 两个子图分别展示了图 1(b) 包含噪声数据集中, 同质样本和异质样本的选取过程. 如图 2(a) 所示, 样本 $A = (\boldsymbol{x}_a, y_a)$ 且有 $y_a =$ 正类, 根据定义 1, 观测标签为正类的样本均为 A 的同质样本, 其中包括两个剩余的假正类样本 B 和 C; 同理, 如图 2(b) 所示, A 的异质样本为观测标签是负类的全部样本, 包括三个假负类样本 D、E 和 F.



(a) 样本 A 的同质样本
(a) Homogeneous sample of sample A



(b) 样本 A 的异质样本
(b) Heterogeneous sample of sample A

图 2　样本 A 的同质、异质样本
Fig. 2　Homogeneous and heterogeneous samples of sample A

定义 2. 绝对离群因子

离群点是数据集中明显异常的数据点. 离群点检测的目的是检测出与正常数据差别较大的数据点. 基于不同的离群点检测方法, 对于数据集中的每个样本, 总能给出一个离群程度的度量, 定义该度量值为绝对离群因子 (Absolute outlier factor, AOF). 根据定义 1, 每个样本点均有相对于其同质、异质样本的一对绝对离群因子.

以 KNN 离群点检测算法为例, 设有二分类数据集 $D=\{(\boldsymbol{x}_i,\, y_i)\}_{i=1}^{n}$, 其中 $\boldsymbol{x}_i\in\mathbf{R}^m$, $y_i\in\{$正类, 负类$\}$. 有样本点 $p=(\boldsymbol{x}_p,\, y_p)$, $q=(\boldsymbol{x}_q,\, y_q)$, 采用欧氏距离度量样本点 p 和 q 间的距离:

$$d\,(p,\, q)=\sqrt{\sum_{t=1}^{m}\left(\boldsymbol{x}_p^t-\boldsymbol{x}_q^t\right)^2} \tag{1}$$

根据给定的距离度量方式, 在 D 中找出与 p 最近的 k 个点, 假定其中距 p 第 k 近的为样本点 q_k, 样本点 p 的绝对离群因子基于其到第 k 近邻样本的距离进行计算:

$$\mathrm{AOF}\,(p,\, D)=k-distance\,(p,\, D)=d\,(p,\, q_k) \tag{2}$$

假定 $y_p=$ 正类, 其同质样本集合记为 D_+, 异质样本集合记为 D_-, 则样本点 p 的同质绝对离群因子定义为:

$$f\,(p)=\mathrm{AOF}\,(p,\, D_+) \tag{3}$$

同理, 样本点 p 的异质绝对离群因子定义为:

$$g\,(p)=\mathrm{AOF}\,(p,\, D_-) \tag{4}$$

基于离群点与噪声点的相似性, 标签噪声往往具有更高的同质绝对离群因子值. 然而, 并不能主观地认为离群因子越高, 其为标签噪声的概率就越大, 即绝对离群因子与噪声概率并不是强正相关关系[22]. 例如标签噪声位于出现概率相近的边界区域, 则该样本点的离群因子值并不会表现得很高; 而离群因子值较高的样本也不一定是标签噪声, 它可能是由于样本本身的特征噪声导致的低概率事件.

以图 1(b) 的人工包含噪声数据集为例, 图 3 四个子图分别给出了基于 KNN 离群点检测算法 (近邻参数 k 设置为 5) 计算部分样本点两类绝对离群因子的结果. 其中, A 为假正类样本, 是噪声样本; G 为真负类样本, 是正常样本.

如图 3(a) 所示, A 点的同质绝对离群因子 $f\,(A)=0.94$, 显然该值相较于其他真正类样本更高; 而图 3(b) 显示其异质绝对离群因子 $g\,(A)=0.36$, 这是由于 A 被大多数与其相反标签的样本点包围, 导致 A 点在异质样本环境中, 具有更低的异质绝对离群因子. 因此, 可以得出含噪样本的同质绝对离群因子往往高于异质绝对离群因子的结论.

在图 3(c) 和图 3(d) 中, 真实样本 G 由于离数据集群较远, 故其两类离群因子都偏高, 特别是其同质绝对离群因子相较于部分噪声样本, 反而更高. 这也印证了绝对离群因子与噪声概率间并不是强正相关关系. 但是, 它的同质绝对离群因子 0.86 低于其异质绝对离群因子 0.65, 该大小关系符合其为正常点的事实. 基于以上分析, 提出相对离群因子的定义来估计样本的噪声概率.

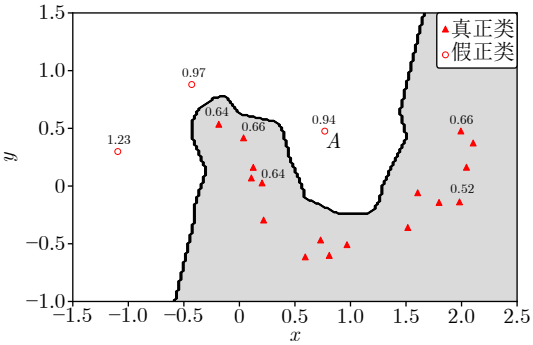
定义 3. 相对离群因子

设有二分类数据集 D, 假定有样本点 $p=(\boldsymbol{x}_p,\, y_p)$, 定义样本 p 的相对离群因子为:

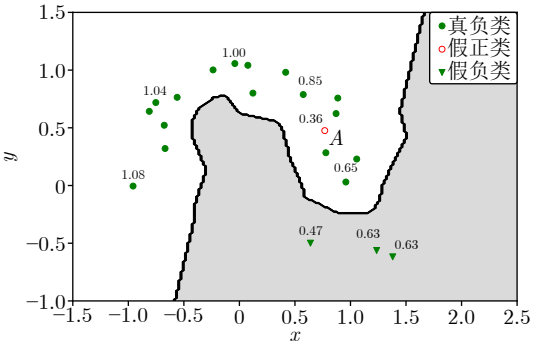
$$\mathrm{ROF}\,(p)=\frac{f\,(p)}{g\,(p)}=\frac{\mathrm{AOF}\,(p,\, D_+)}{\mathrm{AOF}\,(p,\, D_-)} \tag{5}$$

类似地, 若 $y_p=$ 负类, 定义样本 p 的相对离群因子为:

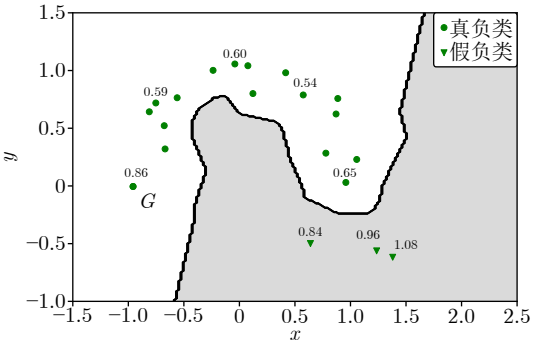
$$\mathrm{ROF}\,(p)=\frac{f\,(p)}{g\,(p)}=\frac{\mathrm{AOF}\,(p,\, D_-)}{\mathrm{AOF}\,(p,\, D_+)} \tag{6}$$



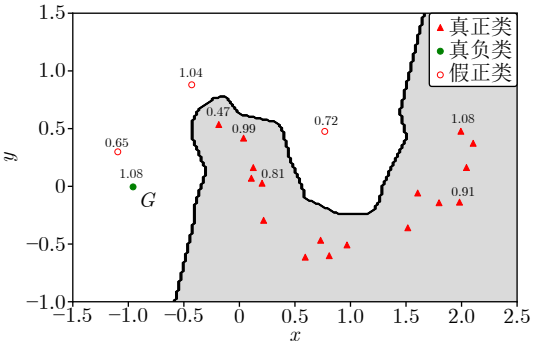
(a) 噪声样本 A 的同质绝对离群因子
(a) Homogeneous absolute outlier factor of noise sample A



(b) 噪声样本 A 的异质绝对离群因子
(b) Heterogeneous absolute outlier factor of noise sample A



(c) 真实样本 G 的同质绝对离群因子
(c) Homogeneous absolute outlier factor of real sample G



(d) 真实样本 G 的异质绝对离群因子
(d) Heterogeneous absolute outlier factor of real sample G

图 3 噪声样本 A 与真实样本 G 的绝对离群因子
Fig. 3 Homogeneous and heterogeneous absolute outlier factor results of noise sample A and real sample G

相对离群因子可用于识别类别型标签噪声, ROF 值越大的样本, 标签噪声的概率也越高. 为确定相对离群因子评估噪声概率的有效性, 沿用 KNN 离群点检测算法 (近邻参数 k 设置为 5), 计算图 1(b) 中数据的相对离群因子值, 部分计算结果如图 4 所示. 其中标签噪声点 (即含噪正类与含噪负类样本) 的相对离群因子分别为 2.49、2.27、2.23、1.79、1.71 和 1.54. 这些值都远大于真实数据点的相对离群因子值. 由图 4 可以看出, 靠近分类边界或远离集群的真实样本, 其同质绝对离群因子值偏大, 导致相对离群因子也偏大, 但相对离群因子值并没有超过噪声样本, 这说明相对离群因子与样本噪声概率的相关关系更强, 在噪声的判别上, 相对离群因子比绝对离群因子更可靠.

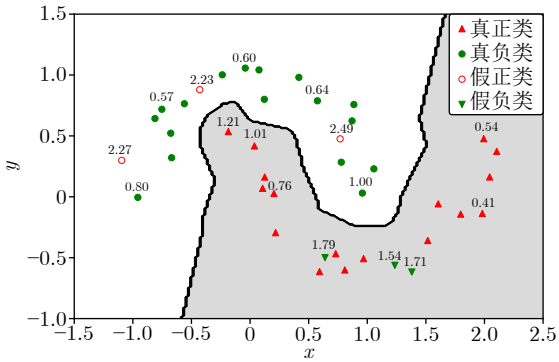


图 4 数据的相对离群因子
Fig. 4 Relative outlier factors for data

总之, 计算样本的相对离群因子能够评估其是噪声的概率, 从而有效检测类别型标签噪声. 据此提出基于相对离群因子的标签噪声过滤算法, 主要步骤如算法 1 所示.

算法 1. 基于相对离群因子的过滤算法

输入. 分类数据集 D, 离群检测模型 p, 参数为过滤阈值 t 或过滤比率 r.

输出. 去噪数据集 D', 噪声集 N.

初始化. 噪声集 $N = \emptyset$.

1) 根据式 (3)、式 (4), 利用离群检测模型 p 计算全部样本的同质、异质绝对离群因子;

2) 根据式 (5)、式 (6), 计算全部样本的相对离群因子;

3) 将数据按相对离群因子大小降序排列;

4) 将相对离群因子大于过滤阈值 t 的样本或前 $n \times r$ 个样本加入噪声集 N 中;

5) 得到去噪数据集 $D' = D - N$.

该算法的时间复杂度主要取决于基离群检测模型计算每个类别离群因子的复杂度. 以 KNN 离群

点检测算法作为基检测模型为例, 若采用 K 维树算法加速搜索 K 近邻[36], 算法 1 的时间复杂度为 $\mathrm{O}(mn\log n)$, 其中 m 为类别数, n 为样本数. 当类别数与样本数均较大时, 类别数对算法的时间开销会显著增加. 但在实际中, 数据集的类别数通常远小于样本数, 因此算法 1 的时间复杂度可近似为 $\mathrm{O}(n\log n)$.

2.2　基于相对离群因子的集成过滤方法

对于不同分布的数据, 基于单一基离群点检测模型计算的相对离群因子无法始终保持噪声概率评估的精确性. 鉴于现有成熟的离群点检测技术, 提出基于相对离群因子的集成过滤方法, 该方法采取贪心策略确定最优联合检测器, 以保证迭代过滤后的数据能训练出最优泛化性能的分类模型, 主要步骤如算法 2 所示.

算法 2. 基于相对离群因子的集成过滤算法

输入. 分类数据集 M, 单次迭代过滤比率 r', 基离群检测器池 $P=\{p_k\}_{k=1}^{m}$.

输出. 去噪数据集 M'.

1) 计算获得基检测器排名池 P':

a) 利用算法 1 (输入. 分类数据集 $D=M$, 离群检测模型 $p=p_k$, 过滤比率 $r=r'$), 获得 m 个不同的去噪数据集;

b) 通过交叉验证, 获得分类模型在不同去噪数据集上的准确率 Acc;

c) 将检测器按对应的准确率降序排列, 得到基检测器排名池 P'.

初始化. 迭代轮数 $k=1$, 缓冲噪声池 $N^*=\emptyset$, 缓冲去噪池 $M^*=M$, 最大准确率 $MAcc=0$.

2) 利用算法 1 (输入. 分类数据集 $D=M$, 离群检测模型 $p=p_k$, 过滤比率 $r=r'$), 得到噪声集 N_k;

3) 同时更新缓冲噪声池 $N^*=N^*+N_k$ 和缓冲去噪池 $M^*=M-N^*$;

4) 通过交叉验证, 获得分类模型在缓冲去噪数据集 D^* 的准确率;

5) 若 $Acc>MAcc$, 则同时更新 $MAcc=Acc$ 和去噪数据集 $M'=M^*$; 否则, 回溯两个缓冲池 $N^*=N^*-N_k$, $M^*=M+N^*$;

6) 若 $k<m$, 令 $k=k+1$, 并重复执行步骤 2) ~ 6); 否则, 停止循环, 获得最终的去噪数据集 M'.

算法 2 的主要流程如图 5 所示, 其中黄色区域表示需要重复计算的步骤, 灰色区域表示缓存的步骤. 算法 2 利用多种互补的基离群点检测算法, 寻求联合检测模型的最优解, 算法主要分为基检测模型排名阶段 (步骤 1)) 和迭代联合过滤阶段 (步骤 2) ~ 6)) 两个阶段. 基检测模型排名阶段检测出的噪声为联合过滤阶段提供多样的去噪组合, 因此在实际运行中, 无需反复调用基检测器进行噪声识别.

该算法的时间复杂度仍主要取决于基离群点检测算法的复杂度, 以第 1.2 节的 6 种基检测器为例, 其中 LOF 算法拥有最高的时间复杂度为 $\mathrm{O}(n^2)$, 在此情况下, 该算法的时间复杂度 $\mathrm{T(EROF)}=\mathrm{O}(n^2)$.

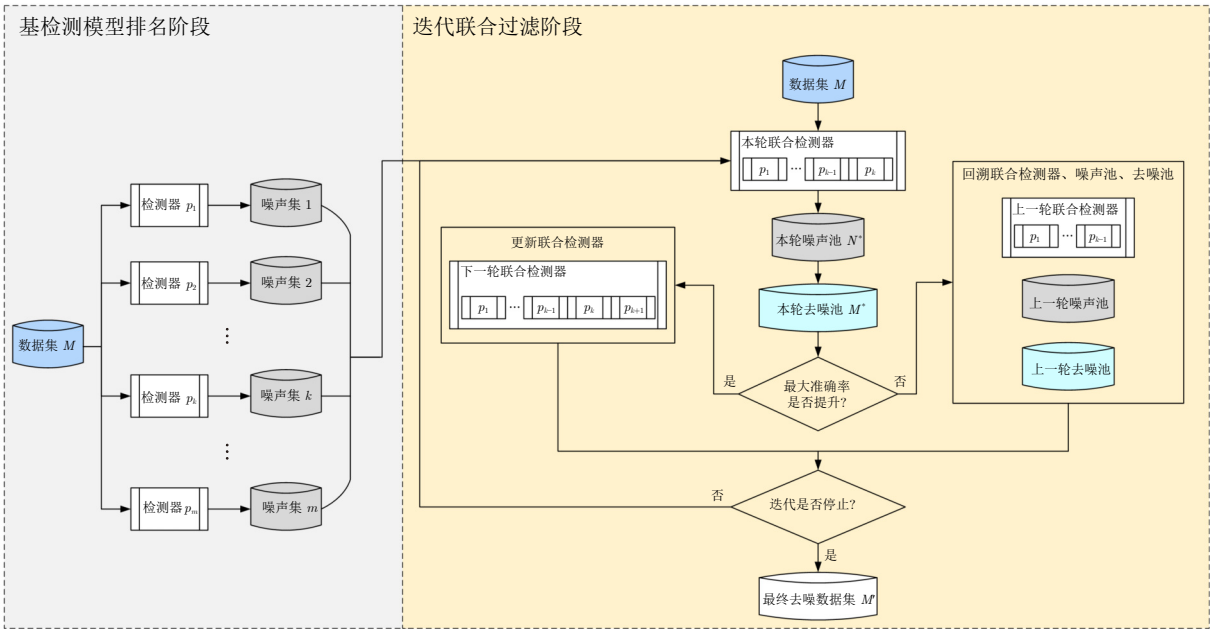


图 5　EROF 算法流程示意图

Fig. 5　Flowchart diagram of the EROF Algorithm

如果基检测器池加入其他模型, 则该算法的时间复杂度与基检测器池中独立检测器的最高时间复杂度相同.

3 UCI 数据集实验

本节介绍基于相对离群因子的集成过滤方法在 UCI 标准数据集上的实验框架、基检测器对比、参数确定、实验结果与相关分析.

3.1 实验框架

实验采用来自 UCI 的 20 个标准数据集, 详细信息见表 1. 这些数据集主要来自真实世界数据, 例如, 第 14 号 Isolet 数据集包含美国英语字母的语音样本, 由不同发音者朗读录制; 第 20 号 Letter 数据集基于 20 种不同字体的黑白矩形像素, 显示 26 个大写英文字母, 特征包括长/宽比、最大水平和垂直笔画数、笔画密度等. 为验证各过滤算法的有效性, 实验将每个数据集按 7:3 的比例, 随机划分为训练集和测试集, 并在训练集上完全随机制造一定比例的人工标签噪声. 首先, 用本文过滤方法与其他主流过滤算法过滤训练集后, 进行各项指标的对比; 然后, 用去噪后的训练集训练分类模型; 最后, 在测试集上, 对比其泛化性能. 由于实验中的人工噪声为完全随机添加, 为降低实验结果的偶然性, 实验结果采用不同随机加噪 10 次的平均值.

表 1 数据集信息

Table 1 Information of data sets

序号	数据集名称	样本数	特征数	类别数
1	Wine	178	13	3
2	Sonar	208	60	2
3	Seeds	210	7	3
4	Heart	270	13	2
5	Votes	435	16	2
6	Musk	476	166	2
7	Wdbc	569	30	2
8	Australian	690	14	2
9	Credit Approval	690	15	2
10	Vehicle	846	18	4
11	Fourclass	862	2	2
12	German	1 000	24	2
13	Svmguide3	1 243	21	6
14	Isolet	1 559	617	26
15	Segment	2 310	18	7
16	Splice	3 175	60	2
17	Satimage	4 435	36	6
18	Banana	5 300	2	2
19	Mushrooms	8 124	22	2
20	Letter	20 000	16	26

为验证算法的有效性, 实验采用准确率、噪声过滤准确率 (Noise filter accuracy, NfAcc)、查准率 (Precision, Pre)、召回率 (Recall, Re)、特异度 (Specificity, Spec) 和 $F1$ 值六种噪声识别评价指标, 分别定义如下:

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

$$NfAcc = \frac{TN}{FN + TN} \tag{8}$$

$$Pre = \frac{TP}{TP + FP} \tag{9}$$

$$Spec = \frac{TN}{TN + FP} \tag{10}$$

$$Re = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{12}$$

式中, 真实结果是正常且预测结果也是正常的样本, 为真正常 (True positive, TP); 真实结果是噪声且预测结果是正常的样本, 为假正常 (False positive, FP); 真实结果是正常且预测结果是噪声的样本, 为假噪声 (False negative, FN); 真实结果和预测结果均是噪声的样本, 为真噪声 (True negative, TN). 在评价指标中, Acc 、$NfAcc$ 、Pre 、Re 和 $F1$ 越高, 表明算法的噪声识别性能越好; $Spec$ 越高, 表明噪声过滤程度越高, 但过高的 $Spec$ 意味着算法可能存在过度清洗问题.

实验设置 10%、20%、30%、40% 四种不同的噪声比例, 选用的六种对比过滤算法及参数设置分别为互近邻过滤器 (近邻参数 k 设置为 3)、多数投票过滤器 (基分类器为 1NN、C4.5 和朴素贝叶斯)、相对密度过滤器 (近邻参数 k 设置为 5, 过滤阈值 rd 设置为 1)、完全随机森林过滤器 (随机树棵数 $Ntree$ 设置为 50, 过滤阈值 $NI\text{-}threshold$ 设置为 5)、基于自适应投票策略的相对密度过滤器 (近邻参数 k 设置为 5) 和自适应完全随机森林过滤器 (随机树棵数 $Ntree$ 设置为 50). 所有过滤算法还与不进行任何过滤 (No filtering, NoF) 的情况进行对比.

最后, 在分类性能对比实验中, 测试所用的分类模型包括 KNNC 模型、支持向量机分类模型、Adaboost 分类模型和随机森林, 分类模型的泛化性能用测试集上的分类准确率来度量.

3.2　基检测器对比

本节实验在 Wine 数据集上, 验证基检测器的多样性, 采用基于 KNN 和 LOF 的两种 ROF 过滤算法, 分别使用 ROF_KNN 和 ROF_LOF 表示算法名称. 噪声比例 NR 和过滤比率 r 均设置为 20%. 通过随机邻域嵌入降维技术, 对数据集进行加噪和去噪的可视化结果如图 6 所示. 其中, 黑色边框标记的样本为未能正确识别出的噪声样本. ROF_KNN 和 ROF_LOF 都实现了良好的噪声过滤效果. 但两个算法在噪声检测上并不完全相同. ROF_KNN 未能过滤的部分噪声样本被 ROF_LOF 成功过滤; 反之, 亦然. 这表明, 多种基检测器间存在互补效应, 结合多种基检测器进行噪声检测和过滤可提高噪声过滤的准确性和可靠性. 因此, 在后续实验中, 本文使用第 1.2 节介绍的 6 种基检测器的组合, 作为 EROF 算法的默认基离群检测器池.

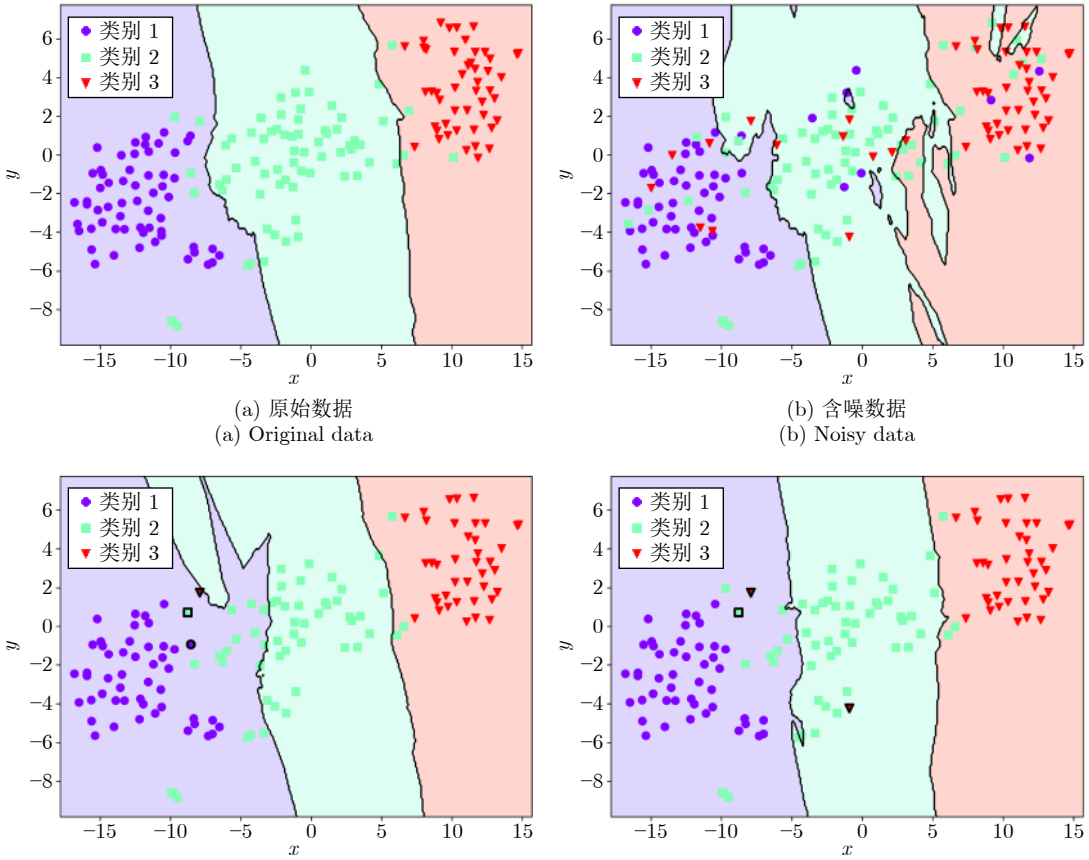
3.3　参数确定

本节实验主要验证过滤比率 r 对 EROF 算法迭代过滤效果的影响. 一般情况下, 算法 2 逐步迭代过滤的过程会使实际过滤噪声的比率超过 r, 因此令 r 在 [0, 0.3] 区间内, 间隔 0.01, 连续取值, 图 7 为在部分数据集上, 加入 4 种不同比例的人工噪声并用 EROF 过滤后, 准确率随过滤比率 r 的变化图, 其中灰色带宽为最优 r 值的集中区域. 可以看出, 当 r 值集中在 [0.04, 0.16] 时, 展现的过滤能力较为优秀; 当 r 值过低时, 过滤样本数量较少, 导致大部分噪声仍未去除, 准确率基本无变化; 当 r 值超过 0.15 后, 由于迭代集成缘故, 在过滤噪声的过程中, 可能去除过多的真实样本, 导致不同噪声比例情况下的准确率都呈现下降趋势. 因此, 在后续实验中, 本文使用 $r = 0.1$ 作为 EROF 算法的默认参数.

3.4　UCI 实验结果与分析

3.4.1　噪声识别性能

图 8 给出了各算法在 4 种噪声比率下, 各项指标的比较结果, 该实验结果取自各算法在 20 个数据集上的平均值.



(a) 原始数据
(a) Original data

(b) 含噪数据
(b) Noisy data

(c) ROF_KNN 过滤
(c) Filtered by ROF_KNN

(d) ROF_LOF 过滤
(d) Filtered by ROF_LOF

图 6　Wine 数据集上, 基检测器噪声过滤对比

Fig. 6　Comparison of base detector noise filtering on Wine

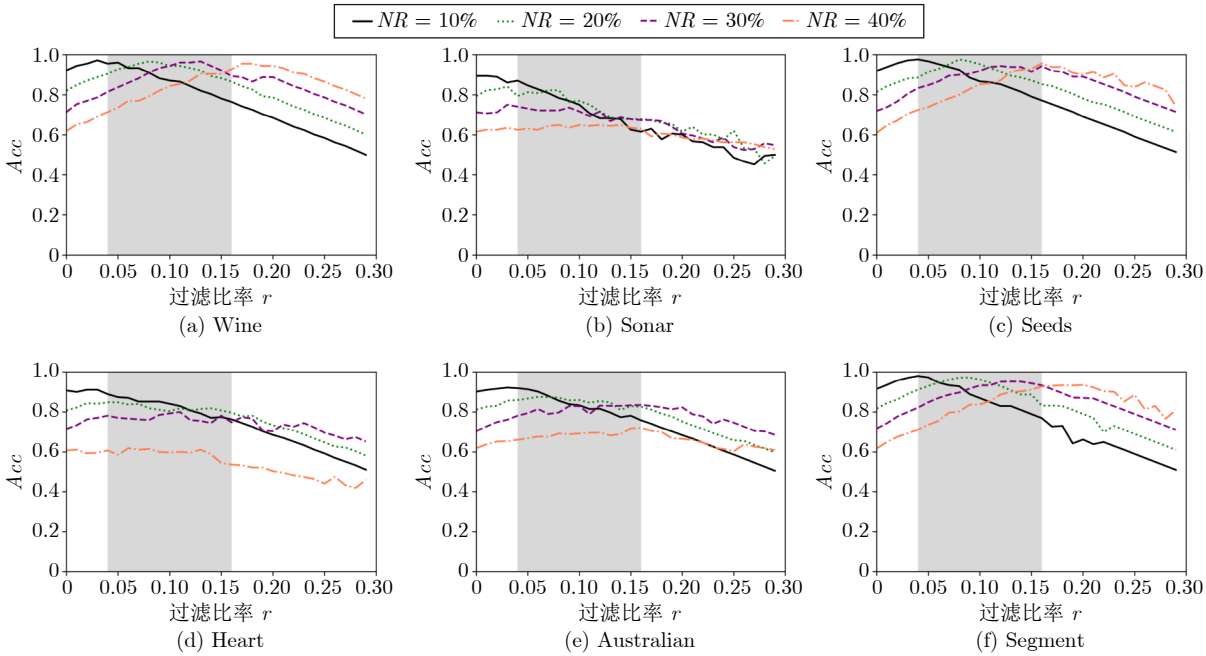


图 7 过滤比率 r 对过滤效果的影响

Fig. 7 Influence of filtering ratio r on filtering effect

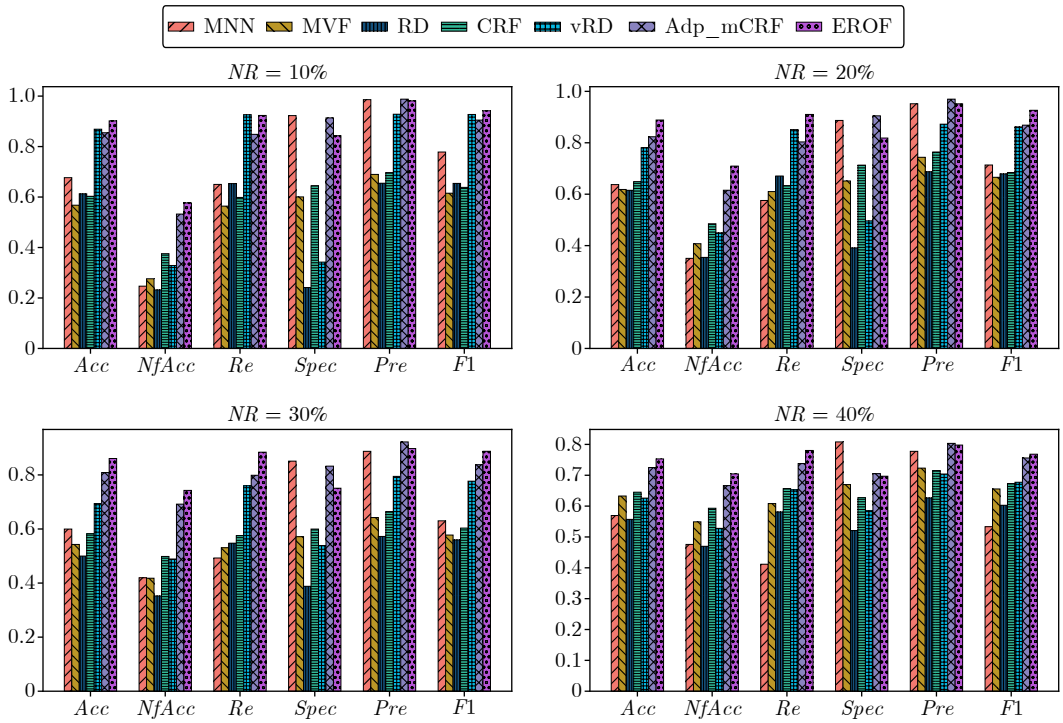


图 8 各算法噪声识别性能指标比较

Fig. 8 Comparison for noise recognition performance indicators of each algorithm

由准确率和噪声过滤准确率可以看出, 当噪声比例在 10% ~ 30% 时, EROF 的噪声识别能力较其他算法有显著优势; 当噪声比例增大至 40% 后, 由于默认设置的 r 值达不到 40% 噪声比例的最优 r 值, 该优势虽有所减弱, 但最终结果仍优于其他算法. 可以看出, 基于分类预测的 MVF 算法在高噪比例下, 很难保持良好的过滤效果, 这是分类模型本身的预测准确性无法保证导致的.

从特异度的角度分析, MNN 和 Adp_mCRF 算法更倾向于尽可能地将噪声去除干净, 因此 *Spec* 表现较为良好. 受聚类效果和迭代停止指标的不确定性影响, vRD 算法的 *Spec* 表现最差. EROF 是以提升分类模型泛化性能为最终目的, 并不总将全部噪声剔除, 而是保留部分对模型影响较小的噪声, 该特性使其 *Spec* 表现相对偏低.

从查准率和召回率看, EROF 的召回率要远高于其他算法, 这是由于其尽可能地保留了真实样本, 也代表其误删真实样本的情况很少发生. 但 EROF 查准率的优势不够稳定, MNN 和 Adp_mCRF 算法在查准率上更占优势. 因此, 由 *Pre* 和 *Re* 两项指标的调和平均 $F1$ 值来对比, 更能体现算法的综合性能. 在 10% ~ 30% 的噪声比例实验中, EROF 的 $F1$ 值稳定保持在 0.9 左右, 而其余算法的 $F1$ 值均接近或低于 0.8. 只有在 40% 噪声比例下, EROF 的 $F1$ 值才跌破 0.8, 这也是由于其在高噪情况下的保守过滤引起的. 总之, EROF 的 $F1$ 值在不同噪声比例的实验中, 都保持着稳定的优势.

图 9 对比了不同噪声比例下, 7 种算法在 20 个数据集中, 各项指标最优次数的占比比例. 在 20% 和 30% 噪声比例实验中, 在除查准率和特异度外的其他评价指标上, EROF 都保持着巨大优势. Adp_mCRF 和 MVF 算法的噪声过滤准确率相较于其他算法更高, 但仍无法超越 EROF 在不同噪声比例下的最优占比.

3.4.2　分类模型泛化性能

表 2 列出了不同噪声比例下, 用 K 近邻分类模型 (近邻参数 k 设置为 1) 在各种算法过滤后的数据集上训练, 然后在无噪测试集上预测的分类准确率结果. 当噪声比例为 10% 时, EROF 在第 2、3、11 号数据集上的分类准确率表现欠佳, 在其余数据集上均为最优, 由于噪声比例偏低, 各算法的分类准确率差距并不明显. 当噪声比例为 20% ~ 30% 时, EROF 在第 1 ~ 10 号和第 16 ~ 20 号数据集上的分类准确率均为最优. 当噪声比例达到 40% 后, EROF 的优势略显不足. 总之, 在所有数据集上, EROF 算法相比于次优算法的分类准确率平均提升了 6.76%, 最大提升了 18.71%; 相比于不过

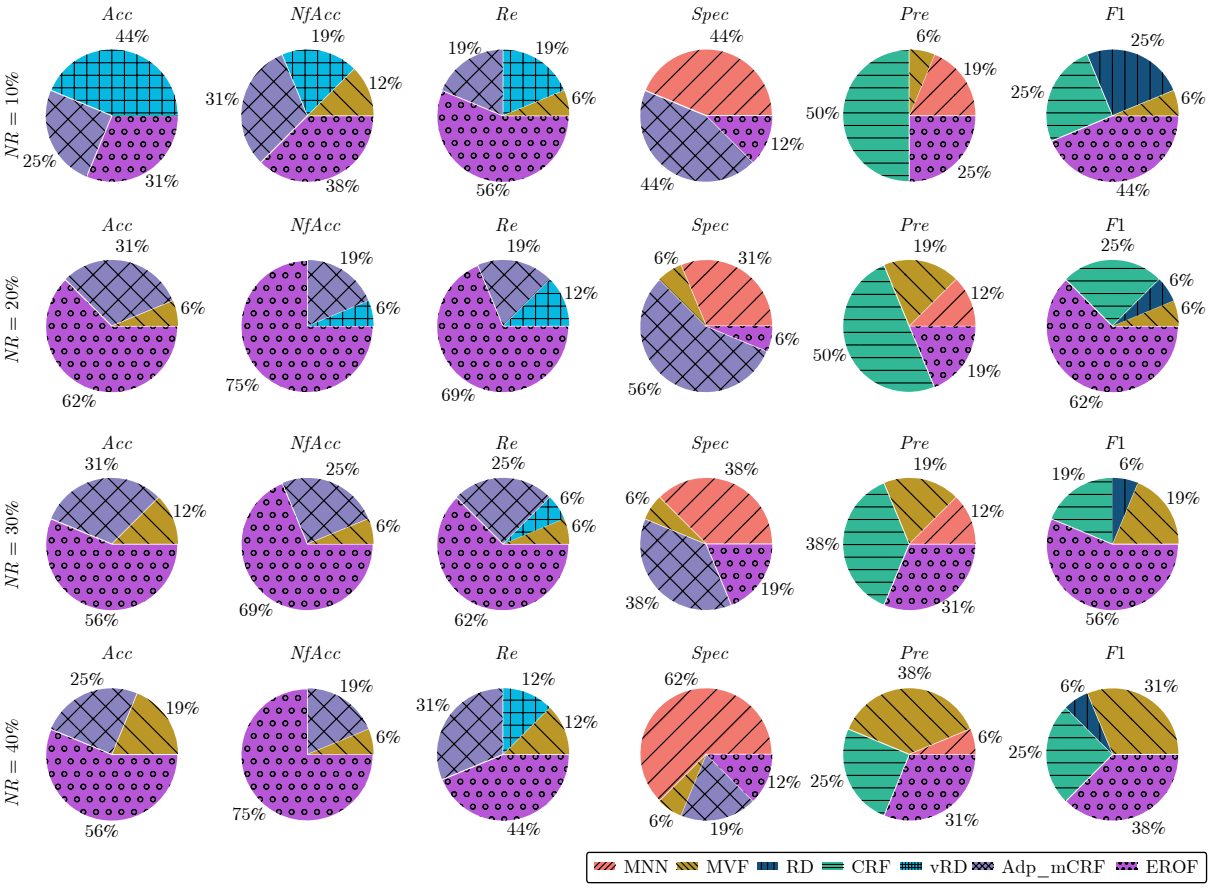


图 9　各算法噪声识别性能指标的最优次数的占比比例

Fig. 9　Optimal frequency ratio for noise recognition performance indicators of each algorithm

表 2 UCI 上, 不同噪声比例下的分类准确率

Table 2 Classification accuracy with different noise ratios on UCI

NR	序号	NoF	MNN	MVF	RD	CRF	vRD	Adp_mCRF	EROF
10%	1	0.926	0.981	0.963	0.921	0.943	0.963	0.963	**0.981**
	2	**0.762**	0.714	0.730	0.713	0.615	0.746	0.635	0.746
	3	0.905	0.921	**0.937**	0.852	0.883	0.889	0.905	0.905
	4	0.691	0.753	0.728	0.665	0.675	0.691	0.704	**0.778**
	5	0.901	0.885	0.901	0.852	0.878	0.885	0.901	**0.908**
	6	0.841	0.841	0.848	0.793	0.807	0.833	0.826	**0.902**
	7	0.727	0.671	0.713	0.687	0.667	0.720	0.699	**0.727**
	8	0.918	0.901	0.912	0.891	0.897	0.936	0.936	**0.959**
	9	0.826	0.855	0.831	0.777	0.785	0.816	0.807	**0.860**
	10	0.841	0.850	0.870	0.804	0.802	0.841	0.831	**0.889**
	11	**0.654**	0.610	0.646	0.630	0.578	0.650	0.591	0.606
	12	0.969	1.000	1.000	0.931	0.933	0.965	0.965	**1.000**
	13	0.697	0.683	0.687	0.681	0.661	0.697	0.683	**0.753**
	14	0.786	0.775	0.786	0.748	0.695	0.772	0.713	**0.812**
	15	0.720	0.726	0.733	0.689	0.676	0.716	0.705	**0.733**
	16	0.931	0.945	0.945	0.889	0.915	0.931	0.938	**0.957**
	17	0.873	0.870	0.879	0.839	0.849	0.872	0.867	**0.893**
	18	0.926	0.936	0.928	0.889	0.884	0.927	0.926	**0.965**
	19	0.972	0.984	0.978	0.942	0.931	0.970	0.969	**1.000**
	20	0.929	0.920	0.941	0.896	0.905	0.921	0.925	**0.949**
20%	1	0.926	0.981	0.981	0.895	0.888	0.926	0.907	**1.000**
	2	0.761	0.762	0.746	0.745	0.669	0.762	0.683	**0.762**
	3	0.778	0.937	0.921	0.753	0.852	0.778	0.873	**0.937**
	4	0.704	0.741	0.704	0.688	0.695	0.704	0.716	**0.753**
	5	0.878	0.794	0.802	0.851	0.836	0.878	0.855	**0.924**
	6	0.818	0.765	0.765	0.795	0.792	0.833	0.826	**0.841**
	7	0.678	0.622	0.650	0.659	0.693	0.685	0.713	**0.715**
	8	0.825	0.830	0.865	0.825	0.780	0.842	0.819	**0.930**
	9	0.758	0.787	0.768	0.746	0.720	0.763	0.749	**0.855**
	10	0.773	0.792	0.739	0.750	0.751	0.773	0.768	**0.845**
	11	0.598	0.563	0.583	0.588	0.536	**0.610**	0.547	0.587
	12	0.942	1.000	0.977	0.911	0.916	0.954	0.938	**1.000**
	13	0.663	0.650	0.663	0.630	0.570	0.663	0.600	**0.720**
	14	0.716	0.681	0.724	0.679	0.658	0.708	0.676	**0.794**
	15	0.662	0.736	0.733	0.641	0.627	0.660	0.652	**0.737**
	16	0.905	0.928	0.935	0.865	0.867	0.908	0.895	**0.958**
	17	0.817	0.823	0.840	0.801	0.781	0.818	0.818	**0.897**
	18	0.854	0.881	0.864	0.830	0.820	0.858	0.862	**0.958**
	19	0.887	0.920	0.930	0.860	0.839	0.882	0.880	**0.998**
	20	0.868	0.855	0.923	0.833	0.829	0.852	0.855	**0.940**
30%	1	0.833	0.944	0.907	0.743	0.686	0.759	0.704	**0.963**
	2	0.603	0.571	0.587	0.580	0.603	0.603	0.619	**0.635**
	3	0.841	0.952	0.921	0.772	0.725	0.794	0.762	**0.957**
	4	0.407	0.420	0.395	0.394	0.580	0.407	0.605	**0.618**
	5	0.710	0.649	0.672	0.652	0.639	0.672	0.664	**0.870**
	6	0.750	0.765	0.795	0.720	0.692	0.750	0.720	**0.818**
	7	0.580	0.594	0.594	0.576	0.597	0.601	**0.622**	0.601
	8	0.684	0.737	0.731	0.666	0.621	0.684	0.649	**0.860**
	9	0.647	0.696	0.696	0.617	0.585	0.643	0.604	**0.807**
	10	0.676	0.647	0.657	0.656	0.649	0.676	0.667	**0.773**
	11	0.531	0.543	0.524	0.502	0.487	0.520	0.512	**0.587**
	12	0.811	0.961	0.927	0.774	0.794	0.815	0.826	**0.973**
	13	0.637	0.633	0.647	0.629	0.586	0.643	0.603	**0.693**
	14	0.651	0.601	0.627	0.627	0.632	0.641	0.649	**0.713**
	15	0.605	0.712	**0.722**	0.541	0.546	0.562	0.562	0.705
	16	0.834	0.922	0.896	0.782	0.777	0.815	0.805	**0.948**
	17	0.740	0.780	0.778	0.717	0.706	0.744	0.739	**0.891**
	18	0.736	0.764	0.747	0.727	0.721	0.744	0.749	**0.924**
	19	0.778	0.870	0.857	0.747	0.752	0.774	0.776	**0.950**
	20	0.791	0.855	0.896	0.747	0.764	0.782	0.784	**0.930**
40%	1	0.630	0.926	0.815	0.588	0.686	0.611	0.704	**0.981**
	2	0.540	0.508	0.508	0.507	0.585	0.524	**0.603**	0.508
	3	0.603	0.794	0.825	0.609	0.590	0.635	0.603	**0.921**
	4	0.457	0.481	0.457	0.458	0.423	0.469	0.444	**0.580**
	5	0.580	0.550	0.511	0.580	0.649	0.603	0.664	**0.718**
	6	0.629	0.674	0.667	0.617	0.591	0.644	0.606	**0.727**
	7	0.573	0.497	0.503	0.561	0.550	0.573	**0.573**	0.524
	8	0.632	0.655	0.661	0.621	0.639	0.643	0.667	**0.766**
	9	0.507	0.531	0.512	0.491	0.490	0.512	0.507	**0.565**
	10	0.551	0.556	0.546	0.547	0.565	0.575	0.589	**0.700**
	11	0.453	0.476	0.472	0.415	0.459	0.429	0.476	**0.555**
	12	0.683	**0.776**	0.757	0.669	0.631	0.699	0.656	0.764
	13	0.580	0.620	**0.627**	0.539	0.551	0.567	0.573	0.577
	14	0.633	0.576	0.582	0.608	0.598	0.622	0.622	**0.641**
	15	0.545	**0.739**	0.720	0.480	0.465	0.494	0.483	0.703
	16	0.688	0.885	0.874	0.677	0.646	0.694	0.680	**0.929**
	17	0.646	0.713	0.707	0.638	0.634	0.653	0.651	**0.884**
	18	0.603	0.618	0.615	0.571	0.593	0.600	0.613	**0.701**
	19	0.652	0.766	0.740	0.614	0.626	0.640	0.645	**0.781**
	20	0.675	0.852	0.857	0.640	0.642	0.666	0.671	**0.922**

滤的分类准确率平均提升了 12.36%, 最大提升了 55.88%, 说明 EROF 算法对 K 近邻分类模型有着较好的增强效果.

图 10 给出了各分类模型分类准确率的临界差异图. 临界差异图能够显示算法排名差异的显著性, 算法排名越小表示算法的分类准确率越高; 算法排名之间的距离不超过临界差异值的用横线连接, 代表算法间的差异不显著. 其中算法的平均排名是基于 20 个数据集和 4 种噪声比例的分类准确率计算得出的. 由图 10 可知, EROF 算法在 4 种分类模型上都取得最优的分类准确率排名. 其中, 在 SVM 和 RF 模型上, EROF 与 Adp_mCRF 算法无显著性差异; 在其他模型上, EROF 算法相较其他算法, 都有显著性优势. 所有过滤算法均优于不进行任何过滤的效果, 这印证了过滤算法能够提升分类模型的泛化性能, 而 EROF 算法的提升效果最高.

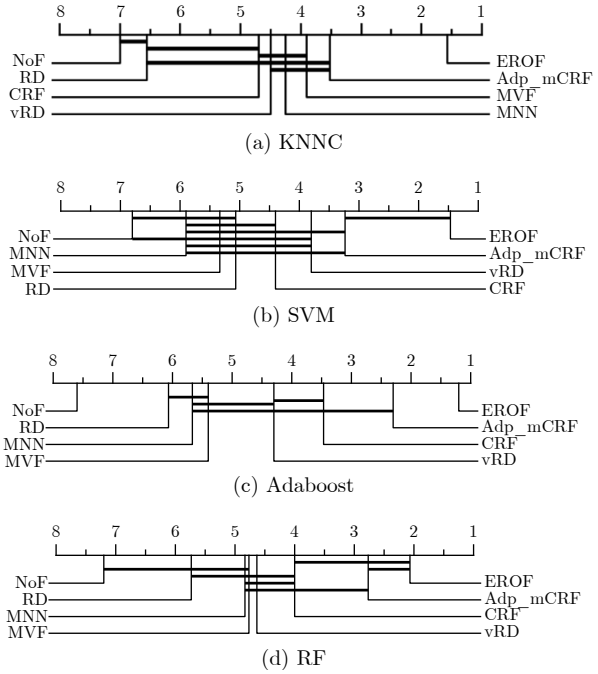


图 10　各分类模型分类准确率的临界差异图

Fig. 10　Critical difference diagram of classification accuracy of each classification model

3.4.3　算法效率

由于部分小规模数据集无法体现算法时间开销差异, 本文选取 Isolet、Mushrooms 和 Letter 三个代表性数据集进行实验, 各算法在这三个数据集上的时间开销对比如图 11 所示. Adp_mCRF 在构建随机树时, 需要随机划分特征进行数据孤立, 导致其在高维数据上效率最低. EROF 算法在多分类任务中, 需要根据类别重复划分同质或异质样本, 并

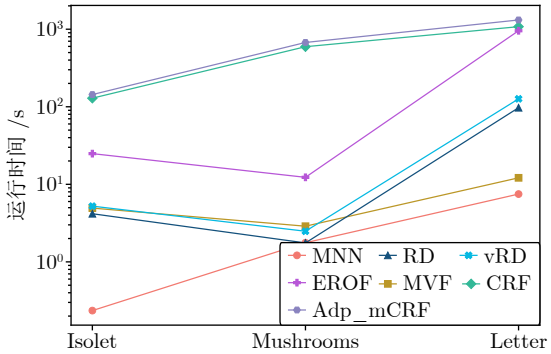


图 11　时间开销对比

Fig. 11　Running time comparison

计算相对离群因子, 因此在有 26 个类别的 Isolet 和 Letter 数据集上, 时间开销较其他算法更高. 与昂贵的模型学习算法和超参数优化算法相比, 优秀的噪声过滤能显著提升模型精度. EROF 算法在保证过滤效果的同时, 相对合理地控制了时间开销, 因此, 其时间开销是可以接受的.

4　MNIST 数据集实验

为进一步验证本文算法对非随机标签噪声识别和过滤的有效性, 本节在 MNIST 标准数据集上进行噪声过滤实验与相关分析.

4.1　实验框架

MNIST 是一个被广泛使用的手写数字图像数据集, 由 60 000 个用于训练和 10 000 个用于测试的 28×28 像素的灰度图像组成. 这些图像表示从 0 到 9 的手写数字, 每个数字大约有 6 000 个图像样本. 在 MNIST 数据集中, 某些数字之间的相似性很高, 它们在书写模糊的情况下, 很难区分, 可能会出现标签噪声的概率也更高. 因此, 本节实验除了在训练集上加入比例为 10% 的完全随机噪声外, 还采用成对翻转的噪声转移矩阵加入非随机标签噪声[27], 其中非常相似类的翻转概率设置为 30%.

实验采用 Adam 优化器, 以 0.001 的学习率迭代训练神经网络 200 次, 其中神经网络具有 2 个隐藏层, 分别包含 500 和 300 个神经元, 可用于处理 784 维输入, 并进行 10 个分类任务[37]. 通过 10 次随机实验, 对不同算法的噪声过滤效果进行对比, 记录 6 种噪声识别性能评估指标和测试集上的分类精度.

4.2　MNIST 实验结果与分析

表 3 列出了在 MNIST 数据集上, 7 种算法对非随机噪声的识别性能. 其中, 在 Acc、$NfAcc$、Re

表 3 MNIST 上的噪声识别性能

Table 3 Noise recognition performance on MNIST

评价指标	MNN	MVF	RD	CRF	vRD	Adp_mCRF	EROF
Acc	0.664	0.732	0.791	0.780	0.801	0.844	**0.872**
$NfAcc$	0.372	0.436	0.489	0.545	0.563	0.651	**0.713**
Re	0.582	0.669	0.739	0.761	0.763	0.816	**0.851**
$Spec$	0.993	0.985	**0.998**	0.858	0.951	0.956	0.957
Pre	0.997	**0.995**	0.999	0.964	0.987	0.988	0.991
$F1$	0.735	0.797	0.850	0.837	0.851	0.885	**0.916**

和 $F1$ 值上, EROF 算法均表现出稳定优势. 尽管 EROF 算法的 $Spec$ 值偏低, 但其优秀的 $NfAcc$ 值表明它能精确地过滤掉数据中的噪声样本. 此外, 0.851 的 Re 值说明, EROF 算法在过滤时优先保留正确样本, 而不是进行无差别的过滤. 这种噪声过滤模式有助于神经网络更好地学习数据的内在规律, 从而提高模型的泛化性能.

图 12 展示了经过不同算法过滤后, 神经网络模型在测试集上的分类准确率, 其中阴影带状区域代表对应算法在 10 次随机实验中的波动范围. 可以看出, 除 MVF 算法外, 其他过滤算法在测试集上的分类准确率均显著优于不进行任何过滤的效果. 其中 EROF、Adp_mCRF 和 vRD 算法的准确率波动较小, 说明高质量的训练数据加速了神经网络拟合新数据的过程.

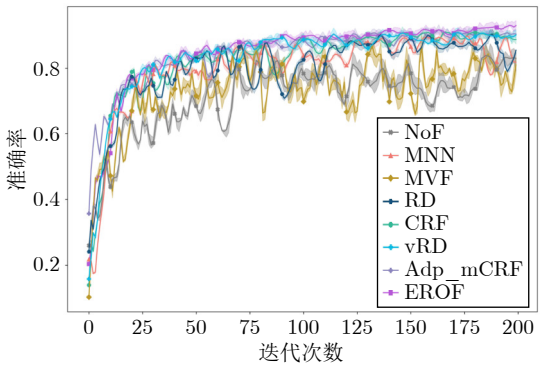


图 12 不同算法过滤后的准确率

Fig. 12 Accuracy after filtering by different algorithms

图 13 给出不同过滤算法在最后 10 次迭代的平均测试精度. 在迭代末期, EROF 算法的平均测试精度达到了 0.925, 为所有算法中的最高值且随机实验对其导致的偏差幅度在所有算法中最小, 表明 EROF 算法具有优秀的精确性和稳定性.

5 结束语

本文提出基于相对离群因子的集成过滤方法,

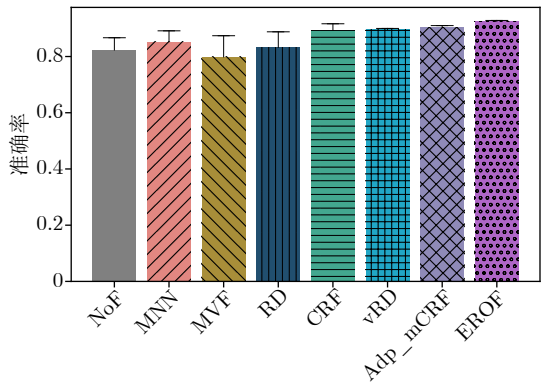


图 13 MNIST 上, 最后 10 次迭代的平均测试精度

Fig. 13 Average accuracy over the last 10 epochs on MNIST

利用基离群点检测算法, 为样本提供标签噪声的概率评估, 再依据此评估实现迭代集成过滤. 与现有算法相比, 该算法以提升分类准确率为最终目的, 在保证数据信息尽量不丢失的同时, 能更精确地过滤掉对分类模型影响较大的噪声样本; 并且该算法通过集成多种互补的离群点检测算法, 保证了其优秀的鲁棒性. 在不同噪声比例和类型下, 该算法均表现出良好的过滤效果和对分类模型的提升能力. 为解决分类任务中的类别型标签噪声过滤问题, 提供一种新的思路.

由于过滤比率 r 和基离群点检测模型对本文算法的过滤效果有一定影响, 因此如何自适应设置过滤比率 r 和选择用于集成的基离群点检测模型, 还有待进一步研究. 此外, 离群点检测算法针对数值型标签噪声问题同样有效, 基于离群点检测算法的相关噪声学习方法在数值型标签噪声过滤和回归模型优化问题上的应用, 值得持续关注与探索.

References

1 Bi Y T, Jeske D R. The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise. *Academic Press*, 2010, **101**(7): 1622−1637

2 Esteva A, Kuprel B, Novoa R A, Ko J, Swetter S M, Thrun S, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 2017, **542**(7639): 115−118

3 Liao Hai-Bin, Xu Bin. Robust face expression recognition based on gender and age factor analysis. *Journal of Computer Research and Development*, 2021, **58**(3): 528−538
(廖海斌, 徐斌. 基于性别和年龄因子分析的鲁棒性人脸表情识别. 计算机研究与发展, 2021, **58**(3): 528−538)

4 Kermany D S, Goldbaum M, Cai W J, Valentim C C S, Liang H Y, Baxter S L, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 2018, **172**(5): 1122−1131

5 Breve F A, Zhao L, Quiles M G. Particle competition and cooperation for semi-supervised learning with label noise. *Neurocomputing*, 2015, **160**: 63−72

6 Zhu X, Wu X. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 2004, **22**(3): 177−210

7 Aversano L, Bernardi M L, Cimitile M, Pecori R, Veltri L. Effective anomaly detection using deep learning in IoT systems. *Wireless Communications and Mobile Computing*, 2021, **2021**: 1−14

8 Khoshgoftaar T M, Van H J. Identifying noisy features with the pairwise attribute noise detection algorithm. *Intelligent Data Analysis*, 2005, **9**(6): 589−602

9 Frenay B, Verleysen M. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**(5): 845−869

10 Yao J C, Wang J J, Tsang I W, Zhang Y, Sun J, Zhang C Q, et al. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 2019, **28**: 1909−1922

11 Jiang Gao-Xia, Wang Wen-Jian. A numerical label noise filtering algorithm for regression. *Journal of Computer Research and Development*, 2022, **59**(8): 1639−1652
(姜高霞, 王文剑. 面向回归任务的数值型标签噪声过滤算法. 计算机研究与发展, 2022, **59**(8): 1639−1652)

12 Sun J W, Zhao F Y, Wang C J, Chen S F. Identifying and correcting mislabeled training instances. In: Proceedings of the Future Generation Communication and Networking. Jeju Isl, South Korea: IEEE, 2007. 244−250

13 Gamberger D, Lavrac N, Groselj C. Experiments with noise filtering in a medical domain. In: Proceedings of the 16th International Conference on Machine Learning. Bled, Slovenia: 1999. 143−151

14 Angelova A, Abu-Mostafam Y, Perona P. Pruning training sets for learning of object categories. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 494−501

15 Brodley C E, Uiversity P, Friedl M A. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 1999, **11**(6): 131−167

16 Sanchez J S, Barandela R, Marques A I, Alejo R, Badenas J. Analysis of new techniques to obtain quality training sets. *Pattern Recognition Letters*, 2003, **24**(7): 1015−1022

17 Sluban B, Gamberger D, Lavrac N. Advances in class noise detection. In: Proceedings of the European Conference on Artificial Intelligence. Lisbon, Portugal: IOS Press, 2010. 1105−1106

18 Tomek I. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems Man and Cybernetics-Systems*, 2007, **6**(6): 448−452

19 Barandela R, Gasca E. Decontamination of training samples for supervised pattern recognition methods. In: Proceedings of the Joint International Workshops of the International Association of Pattern Recognition on Structural and Syntactic Pattern Recognition and Statistical Pattern Recognition. Alicante, Spain: Springer-Verlag, 2000. 621−630

20 Liu H W, Zhang S C. Noisy data elimination using mutual k-nearest neighbor for classification mining. *Journal of Systems and Software*, 2012, **85**(5): 1067−1074

21 Jiang Gao-Xia, Fan Rui-Xuan, Wang Wen-Jian. Label noise filtering via perception of nearest neighbors. *Pattern Recognition and Artificial Intelligence*, 2020, **33**(6): 518−529
(姜高霞, 樊瑞宣, 王文剑. 近邻感知的标签噪声过滤算法. 模式识别与人工智能, 2020, **33**(6): 518−529)

22 Xia S Y, Xiong Z Y, He Y, Li K, Dong L M, Zhang M. Relative density-based classification noise detection. *Optik*, 2014, **125**(22): 6829−6834

23 Xia S Y, Wang G Y, Chen Z Z, Duan Y L, Liu Q. Complete random forest based class noise filtering learning for improving the generalizability of classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 2020, **31**(11): 2063−2078

24 Xia S Y, Chen B Y, Wang G Y, Zheng Y, Gao X B, Giem E, et al. mCRF and mRD: Two classification methods based on a novel multi-class label noise filtering learning framework. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, **33**(7): 2916−2930

25 Huang L, Shao Y, Peng J. An adaptive voting mechanism based on relative density for filtering label noises. In: Proceedings of the IEEE 5th International Conference on Electronics Technology. Chengdu, China: ICET, 2022. 1327−1331

26 Lu J, Zhou Z Y, Leung T, Li J J, Li F F. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: Proceedings of the International Conference on Machine Learning. Stockholm, Sweden: JMLR, 2018. 2304−2313

27 Han B, Yao Q M, Yu X R, Niu G, Xu M, Hu W H, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In: Proceedings of the 32nd Conference on Neural Information Processing Systems. Montreal, Canada: NIPS, 2018. 1−10

28 Xiong H, Pandey G, Steinbach M, Kumar V. Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering*, 2006, **18**(3): 304−319

29 Zhang W N, Tan X Y. Combining outlier detection and reconstruction error minimization for label noise reduction. In: Proceedings of the 6th IEEE International Conference on Big Data and Smart Computing. Kyoto, Japan: IEEE, 2019. 238−241

30 Angiulli F, Pizzuti C. Fast outlier detection in high dimensional spaces. In: Proceedings of the Principles of Data Mining and Knowledge Discovery 6th European Conference. Helsinki, Finland: Springer-Verlag, 2002. 15−27

31 Breunig M, Kriegel H P, Ng R T, Sander J. LOF: Identifying density-based local outliers. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Paris, France: ACM, 2000. 93−104

32 Tang J, Chen Z, Fu A W C, Cheung D W. Enhancing effectiveness of outlier detections for low density patterns. In: Proceedings of the Advances in Knowledge Discovery and Data Mining. Taiwan, China: Springer Berlin, 2002. 535−548

33 Kriegel H P, Schubert M, Zimek A. Angle-based outlier detection in high-dimensional data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA: ACM, 2008. 444−452

34 Schlkopf B, Platt J C, Shawe-Taylor J C, Smola A J, William-

son R C. Estimating the support of a high-dimensional distribution. *Neural Computation*, 2001, **13**(7): 1443−1471

35 Bandaragoda T R, Ting K M, Albrecht D, Liu F T, Zhu Y, Wells J R. Isolation-based anomaly detection using nearest-neighbor ensembles. *Computational Intelligence*, 2018, **34**(4): 968−998

36 Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. In: Proceedings of the International Conference on Management of Data. Dallas, USA: 2000. 427−438

37 Jin Lian-Wen, Zhong Zhuo-Yao, Yang Zhao, Yang Wei-Xin, Xie Ze-Cheng, Sun Jun. Applications of deep learning for handwritten Chinese character recognition: A review. *Acta Automatica Sinica*, 2016, **42**(8): 1125−1141
(金连文, 钟卓耀, 杨钊, 杨维信, 谢泽澄, 孙俊. 深度学习在手写汉字识别中的应用综述. 自动化学报, 2016, **42**(8): 1125−1141)

侯森寓 山西大学计算机与信息技术学院硕士研究生. 主要研究方向为机器学习, 标签噪声.
E-mail: housenyu@163.com
(**HOU Sen-Yu** Master student at the School of Computer and Information Technology, Shanxi University. His research interest covers machine learning and label noise.)

姜高霞 山西大学计算机与信息技术学院副教授. 主要研究方向为机器学习, 数据挖掘.
E-mail: jianggaoxia@sxu.edu.cn
(**JIANG Gao-Xia** Associate professor at the School of Computer and Information Technology, Shanxi University. His research interest covers machine learning and data mining.)

王文剑 山西大学计算机与信息技术学院教授. 主要研究方向为机器学习, 数据挖掘. 本文通信作者.
E-mail: wjwang@sxu.edu.cn
(**WANG Wen-Jian** Professor at the School of Computer and Information Technology, Shanxi University. Her research interest covers machine learning and data mining. Corresponding author of this paper.)