计算机科学
COMPUTER SCIENCE

极限距离噪声估计与过滤方法

姜高霞，秦佩，王文剑

引用本文

姜高霞，秦佩，王文剑.极限距离噪声估计与过滤方法[J]．计算机科学, 2023, 50(6): 151-158.

JIANG Gaoxia, QIN Pei, WANG Wenjian. Noise Estimation and Filtering Methods with Limit Distance [J]. Computer Science, 2023, 50(6): 151-158.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220600130

极限距离噪声估计与过滤方法

姜高霞[1] 秦 佩[1] 王文剑[1,2]
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(山西大学) 太原 030006
(jianggaoxia@sxu.edu.cn)

摘 要 近年来,机器学习不断取得显著性进展并被成功应用于诸多领域,然而很多学习模型或算法高度依赖数据的标签质量。实际应用中大量数据集普遍存在复杂的标签噪声,因此机器学习在低质数据建模和标签噪声处理方面面临严峻挑战。文中针对回归中的数值型标签噪声,从理论分析和仿真实验的角度研究了标签估计区间与噪声的关联性,提出了一种极限距离噪声估计方法。在最优样本选择框架下,基于此噪声估计方法提出了一种极限距离噪声过滤(Limit Distance Noise Filtering,LDNF)算法。实验结果表明,所提噪声估计方法与真实标签噪声具有更高的相关性和更低的估计偏差。在标准数据集和真实年龄估计数据集上证实了所提过滤算法可以在不同噪声环境下有效识别标签噪声并减小模型的测试误差,其表现优于最新的其他过滤算法。

关键词:数值型标签噪声;回归;噪声估计;极限距离噪声过滤;
中图法分类号 TP181

Noise Estimation and Filtering Methods with Limit Distance

JIANG Gaoxia[1],QIN Pei[1] and WANG Wenjian[1,2]
1 School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2 Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education (Shanxi University),Taiyuan 030006,China

Abstract Machine learning has made remarkable progress and has been successfully applied to many fields in recent years. However,many learning models or algorithms are highly dependent on data quality. Complex label noise usually exists in a large number of datasets in practical applications,so machine learning faces severe challenges in low-quality data modeling and label noise processing. To solve the numerical label noise problem in regression,this paper studies the correlation between label estimation interval and the noise from the perspectives of theoretical analysis and simulation experiments,and proposes a limit distance noise estimation method. Under the optimal sample selection framework,a limit distance noise filtering(LDNF) algorithm is proposed based on this noise estimator. Experimental results show that the proposed noise estimation method has a higher correlation and a lower estimation bias with the true label noise. The proposed LDNF algorithm can effectively identify label noises and reduce the test error of the model in different noise environments on benchmark datasets and real-age estimation datasets,and it outperforms other latest filtering algorithms.

Keywords Numerical label noise,Regression,Noise estimation,Limit distance noise filtering

1 引言

作为人工智能的核心技术,机器学习在人脸识别、商品推荐、智慧城市等各个领域有着广泛的应用,这些应用都需要大量高质量数据作为支撑[1-3]。传统机器学习模型通常基于完整、准确、真实的高质量数据进行设计,然而在实际中,数据普遍存在噪声[4-5]。对于监督学习任务,为了降低标注成本,很多实验数据采用众包式标注。然而由于众包平台提供的特征描述不充分[6]以及标注人员认知和专业差异等原因[7],相同样本可能会存在显著的标签差异。通过众包途径得到的数据集虽然庞大且经济但存在严重的噪声问题[7]。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[8]。特征噪声主要指训练样本中属性特征与真实特征之间的差异;标签噪声指训练样本中实际标签与真实标签之间的差异。

到稿日期:2022-06-14 返修日期:2022-11-23
基金项目:国家自然科学基金(U21A20513,62276161,62076154,61906113,U1805263);山西省国际合作重点研发计划(201903D421050)
This work was supported by the National Natural Science Foundation of China(U21A20513,62276161,62076154,61906113,U1805263) and Key R&D Program of Shanxi Province International Cooperation(201903D421050).
通信作者:王文剑(wjwang@sxu.edu.cn)

这两种噪声都会干扰模型的训练和预测，影响模型泛化能力。文献[9-11]的研究表明，标签噪声的危害要远大于特征噪声。标签噪声又可以分为数值型标签噪声和类别型标签噪声。数值型标签噪声的分布复杂性给噪声处理和建模工作带来了极大的挑战，因此设计针对数值型标签噪声的处理方法具有重要的研究意义和价值。

目前，针对标签噪声的处理主要有两个方面。从算法层面，主要是对标签噪声数据的鲁棒建模，这类方法通过调整损失函数、加权、集成等策略设计对标签噪声不敏感的学习算法[12-14]。已有研究结果表明，很多基于鲁棒性建模的方法并没有对标签噪声产生绝对的鲁棒性，模型的泛化能力仍然会受噪声影响[15-16]。从数据层面来看，标签噪声可通过数据清洗过滤和标签纠正来处理[17-19]。标签噪声过滤指直接删除被算法识别为噪声的数据；标签噪声纠正指将错误标签借助专家或算法纠正为真实标签。对比两种方法，标签噪声过滤的成本更低，风险更小。因此，基于标签噪声过滤的方法是处理标签噪声常用的方法。数值型标签噪声问题还可以采取实例选择与过滤算法相结合的处理方式[20]。该方法的主要思路是通过过滤算法将原始数据集中的冗余或者噪声数据删除，在不破坏原始数据集分布的基础上提升数据的质量。数值型标签噪声过滤方法主要分为基于近邻的过滤方法、基于互信息的过滤方法以及基于噪声估计的方法等。

基于近邻的过滤方法是噪声过滤中最常用的一类方法，主要利用近邻模型对标签噪声的敏感性来识别噪声。该类方法最早可追溯到 1968 年，当时 Hart 提出了压缩最近邻算法[21]，它是一种基于 k 近邻规则的实例选择方法。该方法不仅可以识别噪声数据，还可以降低数据集的冗余，但是它的计算结果带有一定随机性。为了降低噪声数据的负面影响，Wilson 等提出了编辑近邻算法(Edited Nearest Neighbor，ENN)[22]。在 ENN 的基础上，All-KNN 算法[23]采用不同近邻值重复筛查噪声数据。Kordos 等将近邻算法用于处理数值型标签噪声问题，提出了编辑近邻回归过滤算法(Edited Nearest Neighbor for Regression，RegENN)[24]，其主要思想是通过比较实例目标值与近邻目标值之间的差异来判断是否存在标签噪声。如果该差异大于某个阈值，就将该实例视为噪声实例并去除，否则保留该实例。Gonzalez 等提出了离散化编辑近邻算法(Edited Nearest Neighbor based on Discretization，DiscENN)[25]，该算法主要是通过数值型标签离散化的方式将原始数据集划分为多个类，将其转化为分类问题再利用 ENN 算法来删除噪声实例。以上算法均属于基于近邻的过滤方法，在一定程度上减少了噪声样本对数据集的影响，但都过于依赖超参数(如近邻数、阈值)的设置，可能会出现过度清洗的问题。

基于互信息的过滤方法(Mutual Information，MI)是在特征选择的启发下提出的。Guillen 等[26]提出在时间序列预测中使用互信息进行实例选择。该方法在去除某个样本和去除其 k 个邻居两种情况下分别计算特征与标签的互信息。如果两种情况下互信息的差异超过设定好的阈值，则将该样本判定为标签噪声样本。当数据集较大时，该方法需要计算大量复杂的互信息，算法效率较低。文献[27-28]提出通过计算训练集中每个实例与当前评估实例之间的互信息，降序排列训练集再与给定阈值对比，从而精简训练集。他们还将这一思想扩展到时间序列预测中的实例选择方法中，减少了计算量，提升了算法效率。

此外，基于近邻的过滤方法和基于互信息的过滤方法都属于启发式过滤，缺乏坚实的理论指导，而且存在超参数设置的问题。

为了解决上述问题，Jiang 等提出了基于噪声估计的覆盖距离过滤(Covering Distance Filtering，CDF)方法[29]。文中从泛化误差界视角提出了最优样本选择框架，解决了噪声过滤的有效性判别和样本自适应选择问题。在此框架下，提出基于噪声估计的 CDF 过滤方法来处理回归和有序分类问题中的标签噪声。CDF 方法为去除标签噪声样本提供了新的思路。文献[30]在 CDF 方法的基础上，通过理论分析噪声水平以及样本量对提升模型泛化能力的影响，提出了相对噪声过滤框架，简化了最优样本选择框架的复杂性，并提出了一种相对噪声过滤(Relative Noise Filtering，RNF)方法，但该方法对低噪声样本存在高估现象。

本文考虑到数值型标签噪声分布的复杂性，针对数值型标签噪声估计不够准确的问题，提出了极限距离标签噪声估计方法；然后在最优样本选择框架下，提出了极限距离标签噪声过滤方法(LDNF)。所提噪声估计方法有效缓解了噪声高估问题，提升了噪声估计的准确度。所提噪声过滤方法提升了数据的质量和模型的泛化能力。

2 相关知识

本节主要介绍相关的基本概念以及最优样本选择框架。

设有回归数据集 $D=\{x_i,y_i\}_{i=1}^n$，其中 x_i 是数据集中第 i 个样本的特征值；y_i 是第 i 个样本的数值型标签。当样本存在标签噪声时，y_i 不等于该样本的真实标签值。令 y_i^0 为第 i 个样本的真实标签值。

2.1 基本定义

定义 1 回归任务中的数值型标签噪声定义如下：

$$e_i=y_i-y_i^0 \tag{1}$$

定义 2 模型 $m(x)$ 在该回归数据集上训练后的模型误差：

$$r_i=m(x_i)-y_i \tag{2}$$

设 D_F 为 D 的过滤数据集，则过滤数据集的样本量 n_F 小于 n。$y=m_F(x)$ 表示在数据集 D_F 上训练得到的模型。

定义 3 将初始数据集与经过过滤后的数据集的过滤比例定义为：

$$\rho=n_F/n \tag{3}$$

定义 4 经验误差定义如下：

$$R_{emp}(m,D)=\frac{1}{n}\sum_{i=1}^{n}[m(x_i)-y_i]^2 \tag{4}$$

由于真实标签值 y_i^0 无法确定，因此常用经验误差值来近似代替真实误差值。

2.2 最优样本选择框架

在最优样本选择框架下，标签噪声过滤的主要目标是以较小的样本损失代价去除掉噪声值较大的样本，从而提高

模型的泛化能力。该框架通过对泛化误差界的等价推导提出了最优样本选择的目标函数。

引理 1[29] 对于回归数据集 $D=\{x_i,y_i\}_{i=1}^n$，当回归模型的拟合优度固定时，该数据集上训练的模型 $m(x)$ 具有最低泛化误差界的一个充要条件为：

$$\min R_{\text{emp}}(m_{\text{F}},D_{\text{F}})\cdot\varepsilon(D_{\text{F}})\Leftrightarrow\max[\beta_T(\rho)-T(\rho)]\cdot\varepsilon(D_{\text{F}}) \tag{5}$$

其中，$E(\cdot)$ 为期望函数，C 为正值系数。

$$\varepsilon(D_{\text{F}})=\varepsilon(h,n\rho,\eta)=(1-\sqrt{[h(\ln(n\rho/h)+1)-\ln\eta]/n\rho})_+^{-1} \tag{6}$$

$$T(\rho)=E_{D_{\text{F}}}(e_i^2)/E_D(e_i^2) \tag{7}$$

$$\beta_T(\rho)=[\varepsilon(D)/\varepsilon(D_{\text{F}})](1+C)-C \tag{8}$$

由引理 1 可知，经过过滤后，模型的泛化误差界主要取决于 $\beta_T(\rho)$ 和 $T(\rho)$ 之间的差值。最优噪声过滤的目标函数为：

$$F(\rho)=[\beta_T(\rho)-T(\rho)]\cdot\varepsilon(D_{\text{F}}) \tag{9}$$

此目标函数所对应的最优过滤比例为：

$$\rho^*=\arg\max_\rho F(\rho)=\arg\max_\rho[\beta_T(\rho)-T(\rho)]\cdot\varepsilon(D_{\text{F}}) \tag{10}$$

由目标函数可知，要想获得最低泛化误差界，$\beta_T(\rho)$ 和 $T(\rho)$ 的差值要最大；而两者中只有 $T(\rho)$ 与 D_{F} 中的噪声有关，因此目标函数需要更小的 $T(\rho)$。保留低噪声样本意味着可以得到一个小的 $T(\rho)$，因此应该优先去除标签噪声大的样本以尽可能降低 $T(\rho)$。

最优样本选择框架提供了用于过滤的最佳样本保留比例，以获得最低的泛化误差界。最优样本过滤比例是综合考虑多种因素的结果，包括样本量、模型误差和噪声水平等。虽然噪声都可以被准确估计，但并不是所有的噪声样本都需要被去除。同时，可以将其他噪声估计方法与此框架相结合生成新的过滤方法，该方法对噪声环境有较好的适应性，可以防止对数据集的过度清洗。

3 基于极限距离的噪声过滤方法

考虑到回归任务中的标签噪声分布比较复杂，已有噪声估计方法不够准确，本文提出了一种新的噪声估计方法，并将其与最优样本选择框架结合，给出了一种新的噪声过滤算法。

3.1 极限距离噪声估计方法

虽然回归数据集中的真实标签是未知的，但可以通过模型预测结果去构造一个大概率包含真实标签值 y_i^0 的区间。当样本的标签噪声较小时，样本的标签值到该区间的距离较小或者刚好落在该区间内；当样本的标签噪声较大时，样本的输出标签值到该区间的距离就会较大。基于这种思想，把难以估计的标签噪声值转化为求样本到标签估计区间的距离值，从而降低识别标签噪声的难度。下面介绍如何有效地构造标签估计区间来度量标签噪声。

为提高构造标签估计区间的准确性，考虑采用子集划分法，将数据集随机划分为 K 个子集，每次选择其中一个子集对基线模型进行训练，然后在全部数据集上进行预测。经过 K 次训练测试后每个样本可得到 K 个预测值，选取其中的最值为每个样本构造标签估计区间。

定义 5 标签估计区间的定义如下：

$$[a_i,b_i]=[\min_k m_k(x_i),\max_k m_k(x_i)] \tag{11}$$

引理 2 若基模型相互独立，则

$$P\{y_i^0<a_i\}=P\{y_i^0>b_i\}=2^{-K}$$

证明：

$$P\{y_i^0<a_i\}=P\{y_i^0<\min_k m_k(x_i)\}=P\{y_i^0<m_k(x_i),\forall k\}=\prod_{k=1}^K P\{y_i^0<m_k(x_i)\}$$

根据等同无知原则，在真实标签未知的情况下可设 $P\{y_i^0<m_k(x_i)\}=P\{y_i^0>m_k(x_i)\}=1/2$，则 $P\{y_i^0<a_i\}=\prod_{k=1}^K\frac{1}{2}=2^{-K}$。同理可得 $P\{y_i^0>b_i\}=\prod_{k=1}^K\frac{1}{2}=2^{-K}$。

由于各个子集的数据不同，它们训练出的模型相互独立，因此上述结论成立。当子集数 K 为大于 1 的正整数时，概率 $P\{y_i^0<a_i\}=P\{y_i^0>b_i\}=2^{-K}$ 较小；而 $P\{y_i^0\in[a_i,b_i]\}=1-2^{1-K}$ 较大，说明式(11)所构造的区间能够以较大概率包含真实标签。但若 K 值过大，子集规模变小，模型误差会变大；K 值过小，标签估计区间极有可能不能包含真实标签，无法进行准确估计。文献[29]的实验结果表明，当 $K=5$ 时实验效果最好，因此本文采取同样设置。对于这种中间概率大两边概率小的分布，可以假设 y_i^0 服从以区间 $[a_i,b_i]$ 为中心的对称分布。在此假设下，研究标签噪声与标签距离 $d=|y_i-c|$ 的关系（$c=(a_i+b_i)/2$），有助于更准确地估计标签噪声。具体的理论关系如定理 1 所示。

定理 1 假设真实标签 y_i^0 服从关于区间中心 c 对称的分布 $f(y)$，即 $f(c-y)=f(c+y)$，其中 $c=(a_i+b_i)/2$，$d=|y_i-c|$，则：

$$\lim_{d\to+\infty}\frac{\partial E|e_i|}{\partial d}=1 \tag{12}$$

证明：期望绝对噪声

$$E|e_i|=\int_{-\infty}^{+\infty}|y_i-y|\cdot f(y)\mathrm{d}y$$

不妨设 $y_i<c$，由 $d=|y_i-c|$ 得 $y_i=c-d$，$d=c-y_i$。令 $t=y-c$ 则有：

$$\begin{aligned}E|e_i|=&\int_{-\infty}^{y_i}(y_i-y)f(y)\mathrm{d}y+\int_{y_i}^{c}(y-y_i)f(y)\mathrm{d}y+\int_{c}^{c+d}(y-y_i)f(y)\mathrm{d}y+\int_{c+d}^{+\infty}(y-y_i)f(y)\mathrm{d}y\\=&\int_{-\infty}^{-d}(y_i-t-c)f(t+c)\mathrm{d}t+\int_{-d}^{d}(c+t-y_i)f(t+c)\mathrm{d}t+\int_{0}^{d}(c+t-y_i)f(t+c)\mathrm{d}t+\int_{d}^{+\infty}(t+c-y_i)f(t+c)\mathrm{d}t\end{aligned}$$

由对称分布的性质可知

$$\begin{aligned}&\int_{-d}^{0}(c+t-y_i)f(t+c)\mathrm{d}t+\int_{0}^{d}(c+t-y_i)f(t+c)\mathrm{d}t\\&\overset{\text{令 }t'=t}{=\!=\!=}\int_{0}^{d}(c-t'-y_i)f(-t'+c)\mathrm{d}t'+\int_{0}^{d}(c+t-y_i)f(t+c)\mathrm{d}t\\&=\int_{0}^{d}(c-t'-y_i)f(t'+c)\mathrm{d}t'+\int_{0}^{d}(c+t-y_i)f(t+c)\mathrm{d}t\end{aligned}$$

$$=\int_0^d (c-t-y_i)f(t+c)\mathrm{d}t+\int_0^d (c+t-y_i)f(t+c)\mathrm{d}t$$
$$=2\int_0^d (c-y_i)f(t+c)\mathrm{d}t$$
$$=2d\int_0^d f(t+c)\mathrm{d}t$$

$$\int_{-\infty}^{-d} (y_i-t-c)f(t+c)\mathrm{d}t+\int_d^{+\infty} (t+c-y_i)f(t+c)\mathrm{d}t$$
$$\underline{\underline{令\ t'=-t}}\int_d^{+\infty} (y_i+t'-c)f(-t'+c)\mathrm{d}t+\int_d^{+\infty} (t+c-y_i)f(t+c)\mathrm{d}t$$
$$=\int_d^{+\infty} (y_i+t'-c)f(t'+c)\mathrm{d}t+\int_d^{+\infty} (t+c-y_i)f(t+c)\mathrm{d}t$$
$$=\int_d^{+\infty} (y_i+t-c)f(t+c)\mathrm{d}t+\int_d^{+\infty} (t+c-y_i)f(t+c)\mathrm{d}t$$
$$=2\int_d^{+\infty} tf(t+c)\mathrm{d}t$$

因此有

$$E|e_i|=2d\int_0^d f(t+c)\mathrm{d}t+2\int_d^{+\infty} tf(t+c)\mathrm{d}t$$

根据函数求导法则可得

$$\frac{\partial E|e_i|}{\partial d}=2\int_0^d f(t+c)\mathrm{d}t+2\mathrm{d}f(d+c)-2\mathrm{d}f(d+c)$$
$$=2\int_0^d f(t+c)\mathrm{d}t$$
$$\underline{令\ y=t+c}\ \ 2\int_c^{c+d} f(y)\mathrm{d}y$$

两边取极限可得

$$\lim_{d\to+\infty}\frac{\partial E|e_i|}{\partial d}=\lim_{d\to+\infty}2\int_c^{c+d} f(y)\mathrm{d}y=2\int_c^{+\infty} f(y)\mathrm{d}y=1$$

定理 1 的理论分析结果表明，期望绝对噪声对标签距离 d 的偏导数趋向于 1，因此当 d 较大时，标签距离 d 可以作为绝对噪声的一种良好估计。在噪声估计的基础上，借助过滤框架可以建立新的噪声过滤算法。可见定理 1 对噪声估计和过滤有直接的理论指导作用。

下面通过仿真实验验证定理 1 的理论结果。如果真实标签服从正态分布，即 $y_i^0\sim N(c,\sigma^2)$，则 $c-y_i^0\sim N(0,\sigma^2)$。由式(1)可得 $e_i=(y_i-c)-(y_i^0-c)=d+c-y_i^0$，因此 $e_i\sim N(d,\sigma^2)$。为了获得稳定的仿真结果，真实噪声 e_i 分别取总体分布 $N(d,\sigma^2)$的0.001∶0.001∶0.999 分位数，然后计算噪声的平均绝对值，作为 $E|e_i|$的结果。在给定方差 σ^2 时，每个 d 值都对应一个期望(平均)绝对噪声 $E|e_i|$。

图 1 给出了不同方差下($\sigma^2=1,2,\cdots,5$)期望绝对噪声 $E|e_i|$与 d 的仿真结果。

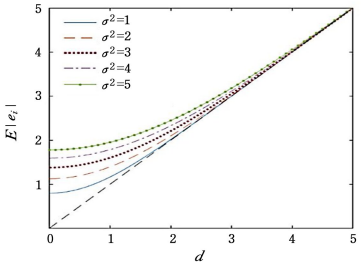


图 1 期望绝对噪声 $E|e_i|$与 d 的仿真结果

Fig. 1 Simulation results of expected absolute noise $E|e_i|$ and d

由图 1 可知，在不同方差情况下期望绝对噪声 $E|e_i|$与 d 都逐渐趋向于 $E|e_i|=d$ 的虚线。当方差较小时，这种趋向速度更快，因此 d 可以作为绝对噪声的一种有效估计。

定义 6 标签噪声的极限距离估计定义如下：

$$LD_i=|y_i-c_i| \tag{13}$$

其中，式(11)区间的中心 $c_i=(a_i+b_i)/2$。

3.2 极限距离过滤方法

根据式(13)，可对数据集中所有样本的标签噪声做出估计，结合最优样本选择框架可实现新的噪声过滤。具体地，首先将标签噪声按照估计值进行降序排列，通过式(9)和式(10)计算目标函数值并得到数据集的最优过滤比例；然后按照标签噪声估计值的大小顺序，优先去除噪声大的样本，保留特定比例的无噪声和低噪声的样本。具体过滤过程如算法 1 所示。其中 step1－step3 对数据集的每个样本构造标签估计区间，计算标签噪声的估计值并将其降序排列；step4－step7 利用 for 循环计算每个样本对应的目标函数，返回目标函数的最大值；step8－step9 由目标函数最大值对应的最优过滤比例，得到过滤后的数据集 D_F。

算法 1 极限距离标签噪声过滤(LDNF)方法

输入：回归数据集 $D=\{x_i,y_i\}_{i=1}^n$；基模型 m(x)

输出：过滤后的数据集 D_F

Step1 将原始数据集 D 随机划分为 5 份，将各个子集轮流作为训练集并将原始数据集作为测试集进行回归预测，得到预测值集合$\{m_k(x_i),i=1,\cdots,n,k=1,\cdots,5\}$；

Step2 利用式(11)为数据集 D 中的样本构造标签估计区间，并根据式(13)计算标签噪声估计值；

Step3 将样本噪声估计值按降序排列，得到新的数据集 D′；

Step4 for j=1 to n

Step5 $n_F=n-j$，$\rho=n_F/n$；

Step6 根据式(6)－式(8)分别计算 $T(\rho)$，$\beta_T(\rho)$，$\varepsilon(D_F)$的值，代入式(9)求得目标函数值；

Step7 end for

Step8 根据式(10)计算数据集的最优过滤比例 ρ^*；

Step9 按照最优过滤比例 ρ^* 在数据集 D′中按噪声估计值进行过滤，得到最优过滤集 D_F。

算法 1 中的 LDNF 方法借助极限距离噪声估计和最优样本选择框架完成对数据集的过滤。实际中基模型采用对噪声敏感的 3 近邻回归模型，以便实现噪声估计。该算法中，除第一步外的其余部分均具有线性时间复杂度，因此算法的时间复杂度与近邻模型相同，均为 $O(n\log n)$。选取的对比算法中 RNF 和 CDF 过滤算法的时间复杂度为 $O(n\log n)$；MI 过滤方法的时间复杂度是 $O(n^3)$；RegENN 和 DiscENN 过滤算法的时间复杂度均为 $O(n^2\log n)$。

4 实验及结果分析

本节主要对所提噪声估计和过滤方法进行实验验证。实验在标准回归数据集上与目前过滤效果较好的方法进行对比，并在真实的年龄数据集上进行验证。

4.1 实验设计

表 1 列出了实验使用的 10 个 UCI 回归数据集的信息[31]。数据均进行归一化处理，以消除奇异样本数据导致的不良影响。

表 1 回归数据集信息

Table 1 Regression dataset information

No	Dataset	# Samples	# Features
1	Forest Fires	517	13
2	Energy Efficiency	768	8
3	Geographical Original of Music	1059	68
4	Airfoil Self Noise	1503	6
5	Skill Craft1 Master Table	3395	20
6	Abalone	4177	8
7	Parkinsons Telemonitoring	5875	26
8	Cpusmall	8192	12
9	Condition Based Maintenance	11934	16
10	Physicochemical Properties of Protein	45730	9

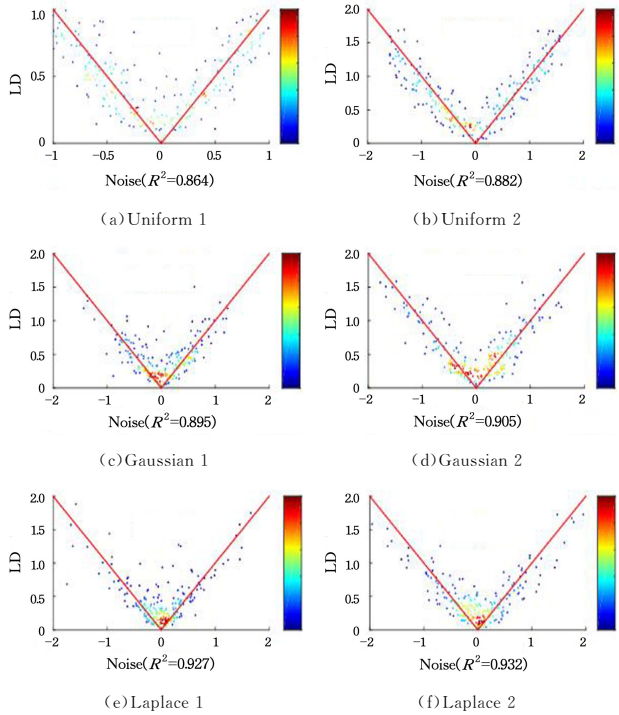
实验过程如下：首先将原始数据集按照 7∶3 的比例随机划分为训练集和测试集，给训练集人工添加多种类型标签噪声形成噪声数据集；然后采用不同过滤方法对噪声数据集进行标签噪声过滤；最后在过滤后的数据集上训练模型，在无噪测试集上测试模型的泛化能力。为了让实验结果更加稳定，每轮实验均重复 5 次。

为了验证本文过滤方法在不同噪声环境下的过滤效果，实验选取 3 种噪声比例，分别为 20%，30%，40%，并选取 6 种标签噪声，它们分别服从均匀分布 1：U(−0.5, 0.5)，均匀分布 2：U(−1, 1)，高斯分布 1：N(μ=0, σ=0.5)，高斯分布 2：N(μ=0，σ=1)，拉普拉斯分布 1：Lp(μ=0，σ=0.5)，拉普拉斯分布 2：Lp(μ=0，σ=1)。此外，选取了 5 种过滤方法：NoF(未过滤)、RegENN[24] 方法(阈值=6，近邻数=9)、DiscENN[25] 方法(近邻数=9)、CDF[29] 方法、RNF[30] 方法，以及本文提出的 LDNF 方法。由于在多个实验结果中其他方法的表现均优于 MI，且 MI 过滤效率较低，因此对比实验中不包括这种过滤算法。

回归测试模型包括 K 近邻(KNN)模型、支持向量机模型(SVR)、高斯过程回归模型(GPR)以及随机森林(RF)。实验结果用均方误差(Mean Square Error，MSE)来度量模型的泛化误差。

4.2 噪声估计实验结果分析

图 2 给出了真实标签噪声与 LD 标签噪声估计的散点图，其中参考线 $LD=|e_i|$ 是理想估计状态，R^2 表示它们的 Pearson 相关系数。考虑到各数据集上的估计性能彼此相似，图 2 给出了数据集 Cpusmall 的结果。由图可见，散点位于红线周围且不同噪声下相关系数均超过 0.85，表明 LD 噪声估计总体上较为准确，且能够区分噪声大小。从噪声分布角度来看，均匀分布 1 中的相关系数小于均匀分布 2 中的相关系数，高斯分布和拉普拉斯分布结果类似。此外，高斯分布 1 的相关系数小于拉普拉斯分布 1，高斯分布 2 的相关系数小于拉普拉斯分布 2。究其原因，主要是均匀/高斯/拉普拉斯分布 1 的方差小于均匀/高斯/拉普拉斯分布 2 的方差，方差较大的分布产生的噪声较大，对应的相关性更加显著。在方差相同的情况下，拉普拉斯分布比高斯分布更广，产生大噪声的概率更大，故相关系数也越大。由此得出结论，LD 噪声估计方法对大噪声估计更准确。



(a) Uniform 1 (b) Uniform 2 (c) Gaussian 1 (d) Gaussian 2 (e) Laplace 1 (f) Laplace 2

图 2 噪声与极限距离的散点图

Fig. 2 Scatter plot of noise and limit distance

表 2 列出了 6 种噪声分布下 CDF，RNF 和 LDNF 过滤方法中的噪声估计与真实噪声的 Pearson 和 Spearman 相关系数以及估计的偏差。噪声估计的偏差采用期望绝对偏差度量：

$$EAD=E\|\hat{|e_i|}-|e_i|\| \tag{14}$$

其中，$\hat{|e_i|}$ 表示对噪声 $|e_i|$ 的估计。相关系数越大表示两者的相关性越强；EAD 值越小表示估计越准确。

由表 2 可知，3 种噪声过滤方法在 6 种噪声分布下噪声估计的 Pearson 值相差无几，Spearman 相关系数中 LDNF 与 RNF 方法相差不大，均略大于 CDF 方法。LDNF 噪声估计的偏差值 EAD 明显小于 CDF 和 RNF 方法的偏差值。这表明 LDNF 噪声估计与真实噪声不仅相关性更大且偏差更小，因此所提 LDNF 方法对噪声的估计更准确。

表 2 6 种噪声与标签噪声估计的相关性

Table 2 Correlation between 6 kinds of noise and label noise estimation

Noise Distribution	Pearson ↑			Spearman ↑			EAD ↓		
	LDNF	CDF	RNF	LDNF	CDF	RNF	LDNF	CDF	RNF
Uniform 1	0.731	**0.732**	0.730	**0.524**	0.521	0.523	**0.183**	0.221	0.194
Uniform 2	0.832	0.833	**0.834**	**0.600**	0.596	0.597	**0.210**	0.268	0.254
Gaussian 1	**0.862**	0.861	0.861	0.555	0.551	**0.556**	**0.232**	0.294	0.253
Gaussian 2	**0.900**	0.897	0.899	**0.596**	0.589	0.594	**0.298**	0.393	0.355
Laplace 1	0.856	0.858	**0.859**	0.477	0.476	**0.478**	**0.240**	0.297	0.247
Laplace 2	**0.893**	0.889	0.887	**0.514**	0.510	0.512	**0.311**	0.394	0.320

4.3 噪声过滤实验结果分析

图3给出了6种过滤方法在3种噪声比例下过滤比例(过滤后保留样本量与原始样本量的比值)的平均值。图中的红线表示理论上的理想过滤比例(1-NR),即恰好将所有噪声样本都去除的过滤比例。当过滤比例高于红线时,保留的噪声样本较多,过滤效果较差;当过滤比例低于红线时,会去除部分非噪声样本,样本损失代价大,效果也较差。因此,过滤比例接近红线时效果最佳。

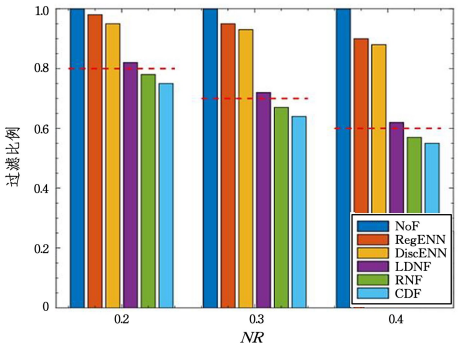


图3 过滤比例对比(电子版为彩图)

Fig. 3 Filtering ratio comparison

由图3可知,每种噪声比例下,RegENN和DiscENN这两种过滤方法对应的过滤比例均值远高于理想过滤比例,过滤后会包含一些噪声样本;CDF和RNF方法的过滤均值均低于噪声比例,LDNF方法的过滤均值略大于噪声比例。这表明RegENN和DiscENN两种过滤方法保留的标签样本过多,效果不佳;CDF和RNF方法过滤比例比前两种更接近理想过滤比例,但可能会去除少量无噪样本;LDNF方法的过滤比例最接近理想过滤比例,它以较小的样本损失代价去除噪声较大的样本,过滤效果最好。

表3列出了不同方法在3种噪声水平下的均方误差值。可以看出噪声比例(NR)越高,对应的均方误差越大。通过横向对比可以看出RNF方法和LDNF方法的误差较小,在高噪声比率下,二者效果相差不大,在噪声比率较低的情况下,LDNF方法的测试误差小于RNF方法。

表3 不同噪声比例(NR)下各模型的均方误差

Table 3 Mean square error of each model with different noise ratios(NR)

NR	Model	NoF	RegENN	DiscENN	CDF	RNF	LDNF
20%	KNN	2.08	1.82	1.77	1.23	1.12	1.02
	SVR	0.82	0.78	0.80	0.70	0.71	0.68
	GPR	0.88	0.82	0.86	0.79	0.73	0.70
	RF	0.97	0.91	0.91	0.72	0.70	0.67
30%	KNN	2.28	2.03	1.94	1.32	1.21	1.18
	SVR	0.87	0.79	0.82	0.78	0.74	0.72
	GPR	0.90	0.95	0.91	0.82	0.79	0.77
	RF	1.05	1.04	1.15	0.80	0.76	0.75
40%	KNN	2.56	2.27	2.16	1.57	1.56	1.56
	SVR	1.08	0.98	0.94	0.87	0.86	0.86
	GPR	1.24	1.07	1.12	0.91	0.89	0.88
	RF	1.45	1.27	1.31	0.92	0.91	0.90

图4给出了4个模型在数据集上的测试误差临界差异图(Critical Difference,CD)。通过对比CD图,不仅可以得到方法优劣的排名,还可以看出方法之间差异的大小,排名越靠前说明误差越小。当各方法之间差异不明显,未超过CD值时用红线连接。

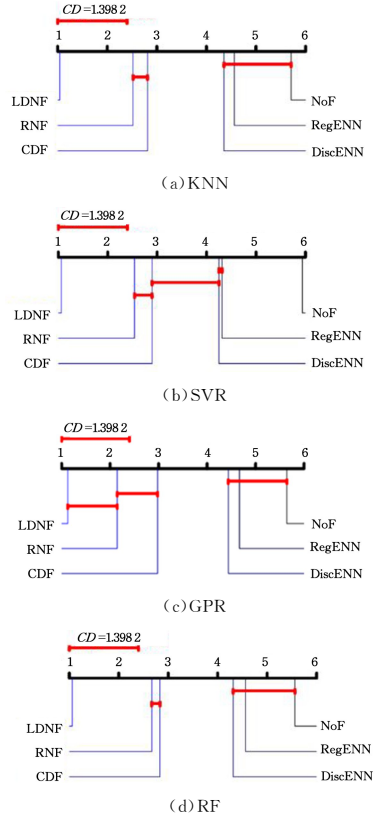


图4 模型测试误差CD图

Fig. 4 CD diagram of model test error

由图4可知,LDNF算法测试误差最小,其次是RNF方法。各种过滤方法均比无过滤方法的误差小很多,这表明针对数据集的过滤方法对模型预测能力有显著提升。

4.4 真实年龄数据集验证

真实年龄数据主要来源于计算机视觉的两大会议ICCV和CVPR的公开年龄标注数据集[32-33],共有18424张图片,每张人脸图片都有对应的年龄估计值(由多名标注人员标注的年龄均值)。在这些数据中存在年龄标记与图片不相符的情况,即存在大量的标签噪声。本文在该数据集上进行标签噪声过滤,并对比不同过滤方法对模型泛化能力的影响。

实验过程设计如下:首先在真实年龄数据集上,用RNF和LDNF两种效果较好的过滤方法进行过滤,得到过滤后的数据集D_F,LDNF方法过滤后大约包括85.5%的原始样本;然后在D_F上分别训练KNN,GPR以及RF这3种回归模型,最后在wiki年龄数据集[33]上进行测试。每轮实验重复5次以获得相对稳定的结果。测试误差使用平均绝对误差(Mean Absolute Error,MAE)作为衡量标准。为了更全面地对比过滤算法的效果,实验将测试样本分为全部测试集、不过滤条件下测试误差大于5和不过滤条件下测试误差大于10的样本集。

实验结果如表4所列,可以看出,在3种情况下LDNF方法与未过滤(NoF)方法相比均可以有效地降低模型测试误差;LDNF方法的模型误差略低于RNF方法,尤其是在MAE大于10的情况下,LDNF方法的效果有明显优势。

表 4 真实年龄数据集上各模型的平均绝对误差

Table 4 Mean absolute error of each model on real-age dataset

Dataset	Model	Samples	Test error		
			NoF	RNF	LDNF
all	KNN	26746	5.40±4.32	5.34±4.30	**5.31±4.28**
	GPR	26746	5.52±4.35	5.48±4.25	**5.45±4.26**
	RF	26746	5.56±4.39	5.54±4.27	**5.53±4.25**
MAE>5	KNN	13625	8.85±4.33	8.80±4.30	**8.78±4.33**
	GPR	13708	8.89±4.28	8.87±4.33	**8.85±4.29**
	RF	13804	9.14±4.41	9.06±4.35	**9.02±4.27**
MAE>10	KNN	3754	13.51±4.46	13.33±4.41	**13.18±4.38**
	GPR	3822	13.55±4.37	13.36±4.34	**13.20±4.26**
	RF	3835	13.82±4.34	13.72±4.32	**13.45±4.46**

实验结果表明，LDNF 方法在真实数据集上可以准确识别并过滤标签噪声，提升了数据的质量和模型泛化能力，是一种针对数值型标签噪声的可行、高效的过滤方法。

图 5 给出了 LDNF 方法识别的部分年龄标签噪声偏差较大的图像。图中给出了图像的名称、原始年龄标签值和经过 5 次重复实验后年龄标签偏差的平均值。其中实线框表示人脸年龄标签偏高（如第 1，2，3，5，9 个），年龄标签偏低的图像用虚线框表示（如第 4，6，7，8，10 个）。LDNF 方法可以准确地找到标签噪声。

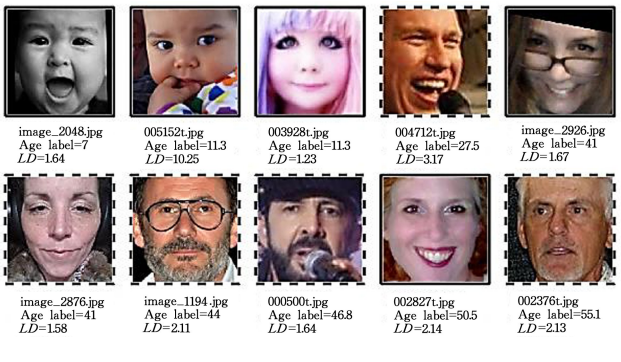


图 5 年龄估计数据集中的部分标签噪声

Fig. 5 Partial label noise in age estimation dataset

结束语 数值型标签噪声问题给回归任务的可靠建模和标签噪声识别与过滤带来了极大困难。本文在研究标签估计区间特征与噪声的关联性基础上，提出了一种极限距离标签噪声估计方法。在最优样本选择框架下，基于此噪声估计方法提出一种高效的数值型标签噪声过滤算法。实验结果表明，所提噪声估计方法与真实标签噪声具有更高的相关性和更低的估计偏差。在标准数据集和真实年龄估计数据集上证实了所提过滤算法可以在不同噪声环境下有效降低各种模型的测试误差，其表现优于最新的其他过滤算法。虽然所提算法是面向回归任务的，但对解决有序回归和分类任务中的标签噪声问题具有一定启发和借鉴意义。此外，如何借助数据纠正和组合策略来提高数据质量，仍值得进一步的研究和探索。

参 考 文 献

[1] ESTEVA，KUPREL B，NOVOA R A，et al. Dermatologist level classification of skin cancer with deep neural networks[J]. Nature，2017，542(7639)：115-118.

[2] MA W J，DONG H B. Face age classification method based on ensemble learning of convolutional neural networks[J]. Computer Science，2018，45(1)：152-156.

[3] KERMANY D S，GOLDBAUM M，CAI W，et al. Identifying medical diagnoses and treatable diseases by image based deep learning[J]. Cell，2018，172(5)：1122-1131.

[4] NORTHCUTTC，JIANG L，CHUANG I. Confident learning：Estimating uncertainty in dataset labels[J]. Journal of Artificial Intelligence Research，2021，70：1373-1411.

[5] KAHNEMAN D，SIBONY O，SUNSTEIN C R. Noise：A flaw in human judgment [M]. New York：Little，Brown Spark，2021.

[6] GUAN D，YUAN W，LEE Y K，et al. Identifying mislabeled training data with the aid of unlabeled data[J]. Applied Intelligence，2011，35(3)：345-358.

[7] MALOSSINI A，BLANZIERI E，NG R T. Detecting potential labeling errors in microarrays by data perturbation[J]. Bioinformatics，2006，22(17)：2114-2121.

[8] ZHU X，WU X. Class noise vs attribute noise：a quantitative study[J]. Artificial Intelligence Review，2004，22(3)：177-210.

[9] LIU G F，ZHAO W Q. Attractors and Their Upper Semi-continuity of Stochastic Lorenz System Driven by Additive Noises [J]. Journal of Chongqing Technology and Business University (Natural Science Edition)，2022，39(1)：78-84.

[10] SAEZ J A，GALAR M，LUENGO J，et al. Analyzing the presence of noise in multi-class problems：alleviating its influence with the One-vs-One decomposition[J]. Knowledge and Information Systems，2014，38(1)：179-206.

[11] FRENAY B，VERLEYSEN M. Classification in thepresence of label noise：a survey[J]. IEEE Transactions on Neural Networks and Learning Systems，2014，25(5)：845-869.

[12] PATRINI G，ROZZA A，MENON A K，et al. Making deep neural networks robust to label noise：a loss correction approach [C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2017：1944-1952.

[13] SABZEVARI M，MARTINEZ-MUNOZ G，SUAREZ A. Vote-boosting ensembles[J]. Pattern Recognition，2018，83：119-133.

[14] 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.

[15] YAO J，WANG J，TSANG I W，et al. Deep learning from noisy image labels with quality embedding[J]. IEEE Transactions on Image Processing，2018，28(4)：1909-1922.

[16] HAN B，YAO Q，YU X，et al. Co-teaching：robust training of deep neural networks with extremely noisy labels [C]// Advances in Neural Information Processing Systems. Cambridge，MA：MIT Press，2018：8536-8546.

[17] CHEN Q Q，WANG W J，JIANG G X. Label noise filtering method based on data distribution [J]. Journal of Tsinghua University(Science and Technology)，2019，59(4)：262-269.

[18] ZHANG Z H，JIANG G X，WANG W J. Label noise filtering method based on local probability sampling[J]. Computer Application，2021，41(1)：67-73.

[19] YU M C，MU J P，CAI J，et al. Noisy label classification learning based on relabeling method[J]. Computer Science，2020，47(6)：79-84.

[20] SEGATA N, BLANZIERI E, DELANY S J, et al. Noise reduction for instance based learning with alocal maximalmargin approach[J]. Journal of Intelligent Information Systems, 2010, 35(2): 301-331.

[21] HART P. The condensed nearest neighbor rule [J]. IEEE Transactions on Information Theory, 1968, 14(3): 515-516.

[22] WILSON D L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Transactions on Systems Man and Cybernetics, 2007, 2(3): 408-421.

[23] CAO J, KWONG S, WANG R. A noise detection based adaboost algorithm for mislabeled data[J]. Pattern Recognition, 2012, 45(12): 4451-4465.

[24] KORDOS M, BIALKA S, BLACHNIK M. Instance selection in logical rule extraction for regression problems [C]// International Conference on Artificial Intelligence and Soft Computing, Berlin: Springer, 2013: 167-175.

[25] ARNAIZ-GONZALEZ A, DIEZ-PASTOR J F, RODRIGUEZ J J, et al. Instance selection for regression by discretization[J]. Expert Systems with Applications, 2016, 54: 340-350.

[26] GUILLEN A, HERRERA L J, RUBIO G, et al. New method for instance or prototype selection using mutual information in time series prediction[J]. Neurocomputing, 2010, 73(10/11/12): 2030-2038.

[27] BOZIC M, STOJANOVIC M, STAJICT Z, et al. Mutual information-based inputs selection for electric load time series forecasting[J]. Entropy, 2013, 15(3): 926-942.

[28] STOJANOVIC M M, BOZIC M M, STANKOVIC M M, et al. A methodology for training set instance selection using mutual information in time series prediction[J]. Neurocomputing, 2014, 141: 236-245.

[29] 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.

[30] 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.

[31] DUA D, GRAFF C. UCI machine learning repository [DB/OL]. [2020-03-28]. http://archive.ics.uci.edu/ml.

[32] HUO Z W, YANG X, XING C, et al. Deep age distribution learning for apparent age estimation[C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2016: 722-729.

[33] ROTHE R, TIMOFTE R, VAN GOOL 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.

JIANG Gaoxia, born in 1987, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include machine learning and data mining.

WANG Wenjian, born in 1968, Ph.D, professor, is a outstanding member of China Computer Federation. Her main research interests include machine learning and computing intelligence.

（责任编辑：何杨）