Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

数据样本的质量评估方法

硕士研究生 马西洋

2025年02月23日

问题回溯

- 总结反思
  - 讲解语速较快，时间较长
  - 内容安排不合理
  - 缺少互动
- 相关内容
  - 2024.01.17 段学明《DNN中的理论可解释性》

内容提要

- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - CG_score
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 了解数据样本质量评估的基本概念及其应用
  - 掌握评估数据样本质量评估中的常用方法和基本概念
  - 理解数据样本在训练模型中的作用

内容引入

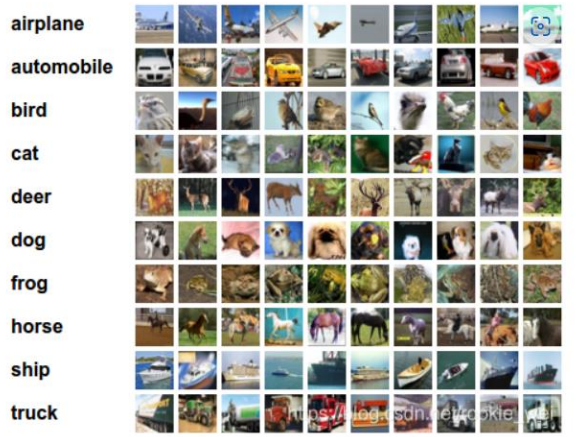
- 著名的二八定律
  - 80%的数据+20%的模型=更好的AI
  - 人工智能是以数据为中心的，而不是以模型为中心
  - 如果我们80%的工作是数据准备，那么确保数据质量是机器学习团队的重要工作
- 为什么数据如此重要?
  - 数据是决策、模型优化、发现趋势和保障研究可靠性的基础，是推动创新和提高效率的关键资源
- 怎么获得更好的数据?

Mr. Krishna, IBM's senior vice president of cloud and cognitive software, said about 80% of the work with an AI project is collecting and preparing data. Some companies aren't prepared for the cost and work associated with that going in, he added.

内涵解析与研究目标

- 题目内涵解析（数据样本的质重评估方法）
  - 数据样本：数据样本是指从总体数据中提取出的一个子集，用于代表和推测整体数据的特征和规律
  - 质量评估：从多个维度来评估数据的准确性、完整性、一致性和重要性等方面，目的是确保数据能够提供有效的支持用于分析或决策
    - 准确性：数据是否正确，是否包含错误或偏差
    - 完整性：数据是否缺失，是否存在缺失值或空值
    - 一致性：数据中是否存在矛盾或冲突的信息，多个数据源是否一致
    - 代表性：数据样本是否能够代表总体样本，是否存在偏倚

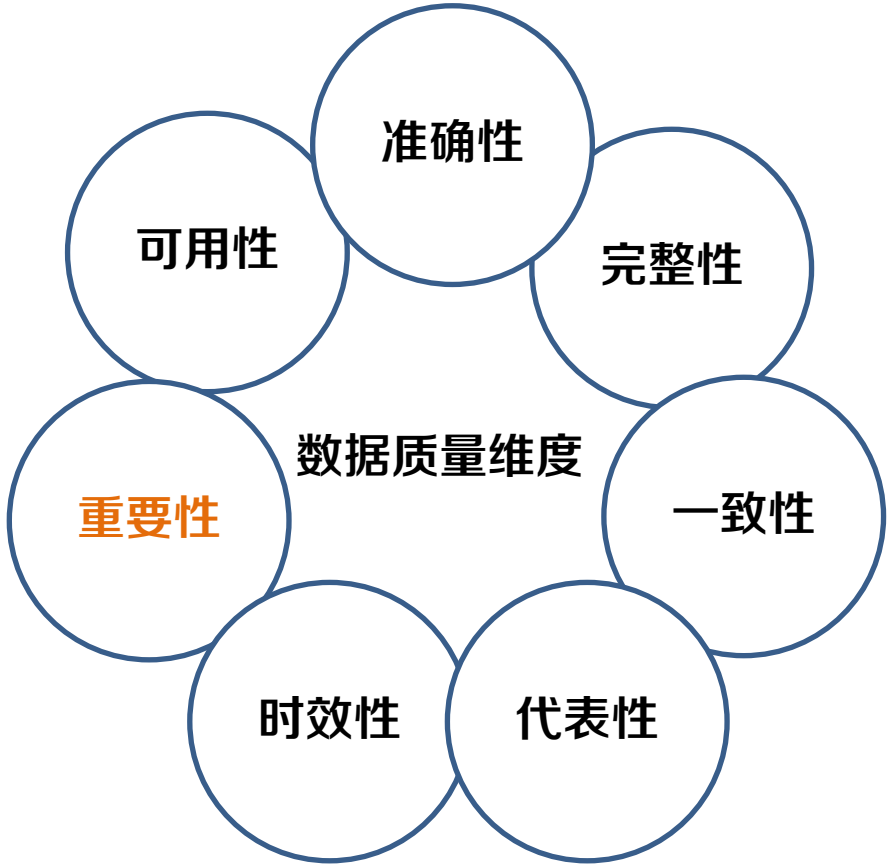
结构化数据



非结构化数据

内涵解析与研究目标

- 题目内涵解析（数据样本的质重评估方法）
  - 质量评估：
    - 时效性：数据是否更新及时，是否存在过时的数据
    - 可用性：数据是否易于访问、理解和处理，是否符合目标需求
    - 重要性：数据是否显著影响模型预测结果
- 研究目标
  - 以数据集为研究对象，面向机器学习任务
  - 结合聚类分析、模型性能评估和数据估值等技术
  - 实现对数据样本在机器学习中的重要程度进行快速且精确的评估

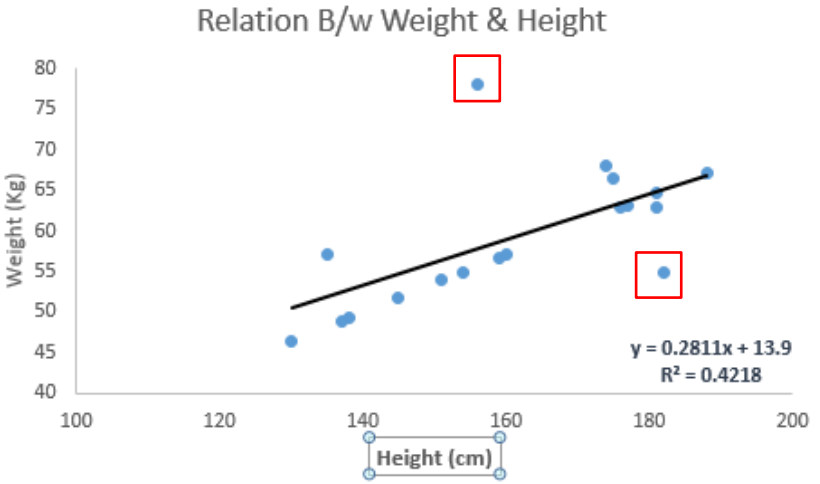


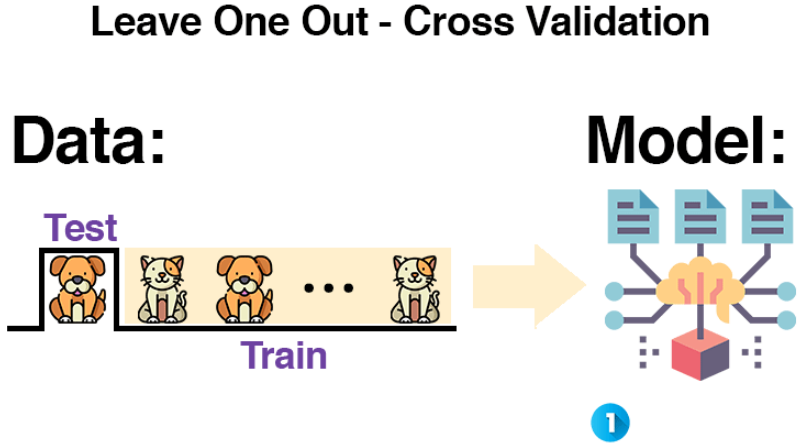
研究背景与意义

- 研究背景
  - 数据质量的问题
    - 现实中的数据往往存在各种质量问题，如数据缺失、错误、不一致等
  - 现有方法局限
    - 传统的数据质量评估方法复杂且计算开销大
- 研究意义
  - 提高模型训练效率：
    - 可以有效去除冗余、噪声或不相关的样本，减少无效样本对训练的干扰
  - 提高数据可解释性：
    - 可以帮助理解哪些数据对模型决策产生了影响，有助于提升模型的可解释性

研究历史　数据样本的质量评估方法

- 传统方法
  - 杠杆值（Leverage Score）
    - 反映了某个样本在自变量空间中的位置，距离其他数据点的远近
    - 具有高杠杆值的数据点可能是异常点或对模型有较大影响的点
    - 只适用于线性回归等回归模型；依赖于特征空间
  - 留一法（Leave-One-Out）
    - 每次将一个样本作为测试集，其余样本作为训练集
    - 优势：
      - 能够提供相对精确的模型性能评估
    - 劣势：
      - 计算开销大，尤其是数据集较大时
      - 可能导致较大的方差，评估结果不稳定





研究历史　数据样本的质量评估方法

- 数据沙普利值（Shapley Value）
  - 用于评估每个训练样本对模型性能（如精度、误差等）的影响

$$\phi_i = \sum_{S \subseteq D-\{i\}} \frac{|S|!\,(|D| - |S| - 1)!}{|D|!} V(S \cup \{i\}) - V(S)$$

  - D是训练集，S是除数据i以外其他训练集的所有子集，$V(S)$表示在数据S上训练的预测器的性能得分

优势	劣势
提供了公平的数据样本重要性评估	计算开销大；计算复杂度高
不依赖于特定的模型结构	常采用近似方法，无法保证绝对准确性
量化每个样本的贡献，提高解释性	多次训练模型可能产生误差

研究历史　数据样本的质量评估方法

- 数据沙普利值（Shapley Value）
  - 共有三家公司1、2、3
  - 公司1，2，3单独投资可盈

$v(1)=100, v(2)=200, v(3)=300$

  - 如果公司1和公司2联合，可获利

$v(1\&2)=500, v(2\&3)=600,$

$v(1\&3)=700$

  - 公司1、公司2和公司3联合，可获利$v(1\&2\&3)=1000$
  - 每个公司各获利多少?
  - 公司2：$\frac{850}{3}$ 公司3：$\frac{1300}{3}$

S	0	2	3	2、3
$V(S\cup\{i\})$	100	500	700	1000
$V(S)$	0	200	300	600
$V(S\cup\{i\})-V(S)$	100	300	400	400
$\lvert S\rvert$	0	1	1	2
$\lvert D\rvert$	3	3	3	3
$\frac{\lvert S\rvert!\,(\lvert D\rvert-\lvert S\rvert-1)!}{\lvert D\rvert}$	$\frac{0!\,2!}{3!}$	$\frac{1!\,1!}{3!}$	$\frac{1!\,1!}{3!}$	$\frac{2!\,0!}{3!}$
Shapley	$\frac{100}{3}$	$\frac{150}{3}$	$\frac{200}{3}$	$\frac{400}{3}$
Sum	$\frac{850}{3}$			

研究历史　数据样本的质量评估方法

Amirata等人提出了数据夏普利值（Data Shapley）作为量化每个训练数据对预测器性能的价值的指标，并开发了蒙特卡洛和基于梯度的方法来高效估算数据夏普利值

2019

2019

Mariya等人以灾难性遗忘为灵感，发现任务中的一些例子更容易被遗忘，而另一些则始终难以忘怀，可以根据模型的遗忘统计识别重要样本、检测异常值和具有噪声标签的示例

Garima等人提出了TracIn方法，通过追踪训练过程中每次使用特定训练样本时，测试点损失的变化，来计算该训练样本对模型预测的影响

2020

2021

Jiang提出通过结合性得分来分析模型如何处理单个样本，该分数衡量了样本是否遵循某种特定的规律或关联模式，具有高度结构化的样本C_score值较高

Kwon 等人针对估算 Data Shapley 的计算成本高，以及关于Data Shapley值如何取决于数据特征的数学分析少的问题，为线性回归等典型问题推导出了Data Shapley的解析表达式

2021

2022

Zhao等人针对计算影响函数的方法很脆弱，且仍然缺乏在神经网络背景下理论分析的问题，利用神经切线核（NTK） 理论计算了用正则化均方损失训练的神经网络的影响函数，使其更适用于高维过参数化模型

Wang等人针对随机梯度下降的固有随机性会导致现有的数据值概念在不同运行中产生不一致的数据值排名的问题，将Banzhaf值应用于数据估值，提高稳健区分数据质量的能力

2023

2024

Hong 等人针对计算 Data Shalpey需要大量对模型的重复训练的问题，提出先确定相似数据点群集的值，再令该值所有成员群集点之间进一步传播，大大降低了计算量

研究现状

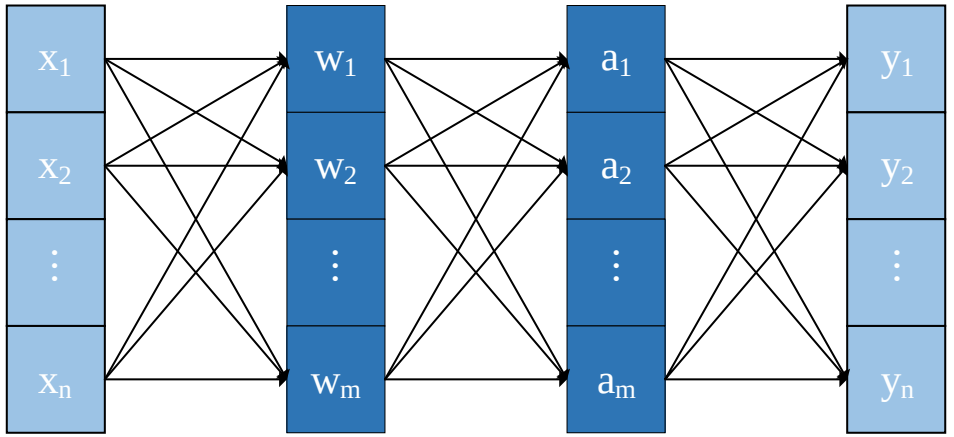
- 减少传统数据沙普利值的资源和时间消耗
  - 近似计算
    - 通过使用蒙特卡罗方法或其他启发式算法，对沙普利值进行近似计算
  - 数据点集评估
    - 通过评估数据点集而非单一数据点，减少对单个样本的过度依赖，提升评估效率
  - 无需实际训练
    - 分析数据样本的复杂度差异，评估其对模型参数的贡献
- 建立多维度、体系化的评价标准
  - 结合性能、稳定性等多维度指标，建立全面的数据重要性评估标准
  - 统一评估框架
    - 避免在不同研究或应用场景中出现评估标准不一致的问题

CG_score: DATA VALUATION WITHOUT TRAINING OF A MODEL

CG_score TIPO

T	目标	更加快速的估算样本重要性
I	输入	图片数据样本 FMNIST（70000张灰度图像，大小28x28，分为10类） CIFAR-10（60000张彩色图像，这些图像是32*32，分为10类） CIFAR100（60000张彩色图像，这些图像是32*32，分为100类）
P	处理	1.由随机初始化梯度下降训练两层神经网络 2.将样本输入训练好的神经网络 3.根据神经网络参数偏移确定评分
O	输出	每个样本的评分（0~1）

P	问题	现有的基于Shapley值的机器学习数据估值框架计算成本高昂
C	条件	需要提前训练较好的神经网络
D	难点	降低计算计算复杂度；提高评分的准确性
L	水平	ICLR 2023（CCF-A）



- 两层神经网络中的数据复杂度度量
  - 由随机初始化梯度下降训练两层神经网络
  - 定义了Gram matrix：
  - $H_{ij}^{\infty} = \mathbb{E}_{\omega \sim \mathcal{N}(0, I_{d*d})}\left[x_i^T x_j 1\{\omega^T x_i \geq 0, \omega^T x_j \geq 0\}\right]$

    $= \frac{x_i^T x_j (\pi - arccos(x_i^T x_j))}{2\pi}$
  - $x_i^T x_j$衡量了点的相似性，$\omega \sim \mathcal{N}(0, I_{d*d})$ 表示 从标准正态分布中随机采样的向量ω，$1\{\omega^T x_i \geq 0, \omega^T x_j \geq 0\}$是指示函数，当且仅当$\omega^T x_i \geq 0, \omega^T x_j \geq 0$时，函数值为1，$\mathbb{E}_{\omega}$表示对$\omega$进行期望计算
  - 目标：在多个随机方向上综合评估数据点之间的相似性，而不仅仅是在一个固定的方向上，够反映出数据在高维空间中的复杂几何结构和非线性关系

CG_score 前置知识

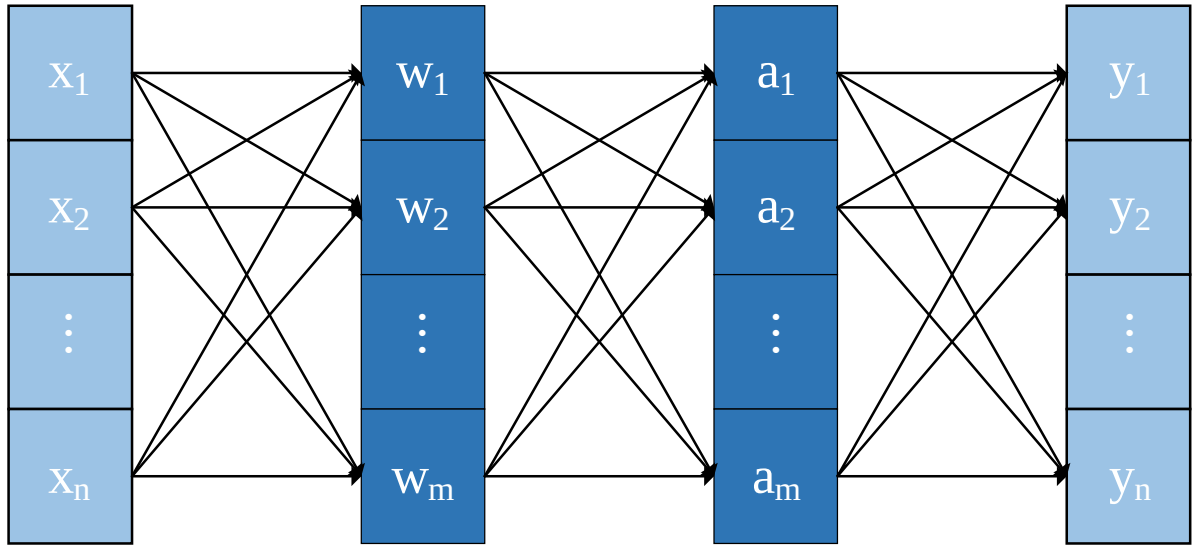
- 两层神经网络中的数据复杂度度量
  - 数据的复杂度度量定义为：

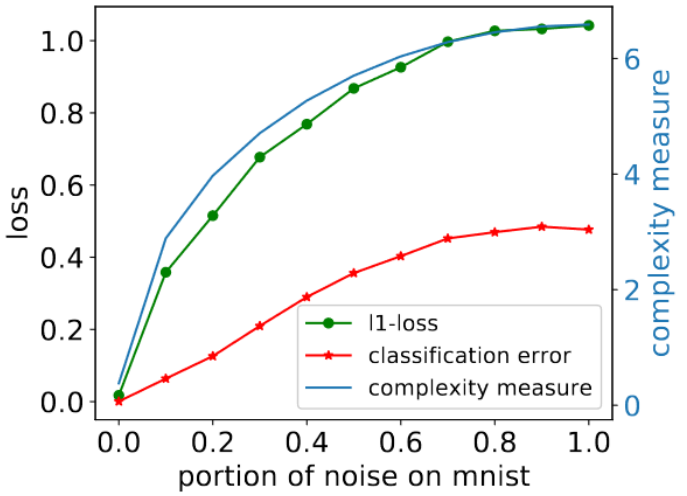
$$y^T(H^\infty)^{-1}y$$

  - 随着随机标签的一部分的增加，复杂度衡量几乎与误差的趋势相匹配
  - 神经元的全部变化之和：

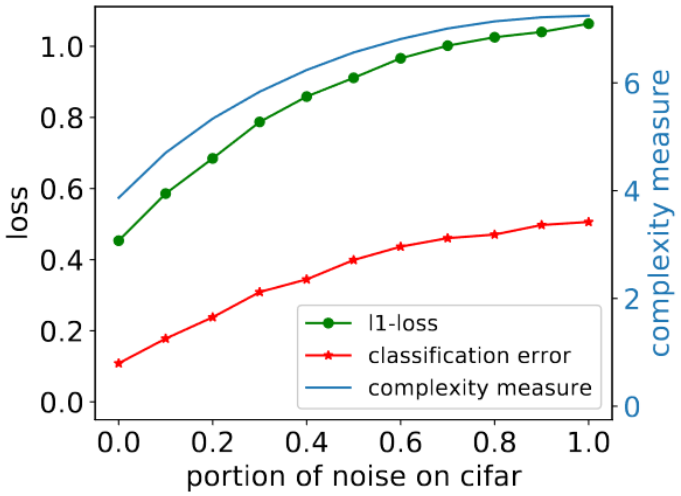
$$||W(k)-W(0)||_F^2 \leq y^T(H^\infty)^{-1}y+\text{small constant}$$

为什么不用H^∞？





(a) MNIST Data.



(b) CIFAR Data.

CG_score 两个重要属性

- 复杂性差距得分（complexity-gap）

$$CG(i) = y^T(H^{\infty})^{-1}y - y_{-i}^T(H_{-i}^{\infty})^{-1}y_{-i}$$

  - 具有较大CG得分的实例（x_i, y_i）是一个“困难”示例
    - 从数据集中将其删除的意义上，将其降低了大量限制的概括误差，这意味着没有（x_i, y_i）的数据集更容易学习
  - 具有较大CG得分的实例（x_i, y_i）在优化方面具有更多的贡献
    - 通过$||W(k) - W(0)||_F^2$来驱动更多的神经元总变化
- 计算复杂度的问题
  - 矩阵求逆的计算复杂度为O（n^3），总的计算复杂度为O（n^4） 难以接受!
  - Schur complement（舒尔补矩阵） 为什么不用别的简化方法?
  - $H^{\infty} = \begin{pmatrix} H_{n-1}^{\infty} & g_i \\ g_i^T & c_i \end{pmatrix}$ $(H^{\infty})^{-1} = \begin{pmatrix} (H^{\infty})_{n-1}^{-1} & h_i \\ h_i^T & d_i \end{pmatrix}$

CG_score 计算复杂度的问题

- 计算复杂度的问题
  - 根据舒尔补矩阵：

我确信已发现了一种美妙的证法，可惜这里空白的地方太小，写不下

$$(H_{-i}^{\infty})^{-1} = (H_{n-1}^{\infty})^{-1} = (H^{\infty})_{n-1}^{-1} - d_i^{-1} h_i h_i^T$$

$$y^T (H^{\infty})^{-1} y = y_{-i}^T (H_{n-1}^{\infty})^{-1} y_{-i} + y_i h_i^T y_{-i} + y_i y_{-i}^T h_i + y_i^2 d_i$$

$$y^T (H_{-i}^{\infty})^{-1} y = y_{-i}^T (H_{n-1}^{\infty})^{-1} y_{-i} - d_i^{-1} (y_{-i}^T h_i)^2$$

$$CG(i) = \left(\frac{y_{-i}^T h_i}{\sqrt{d_i}} + y_i d_i\right)^2$$

$CG(i)$与损失有关！

  - 可以不计算$(H_{-i}^{\infty})^{-1}$得到$CG(i)$

$$\sum_{k=0}^{\infty} \left(y_i - u_i(k)\right) \approx \left(\frac{y_{-i}^T h_i}{\sqrt{d_i}} + y_i d_i\right) \sqrt{d_i}/\eta$$

CG_score 实验设计

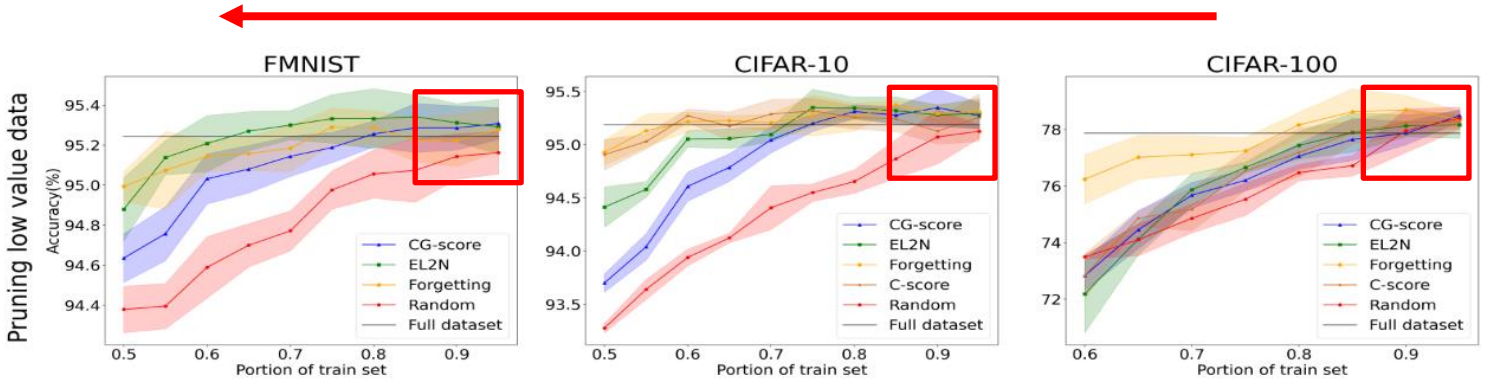
- 数据资源
  - 数据集：

数据集	类别数	每个类别的样本数	样本总数
FMNIST	10	7000	70000
CIFAR-10	10	6000	60000
CIFAR-100	100	600	60000

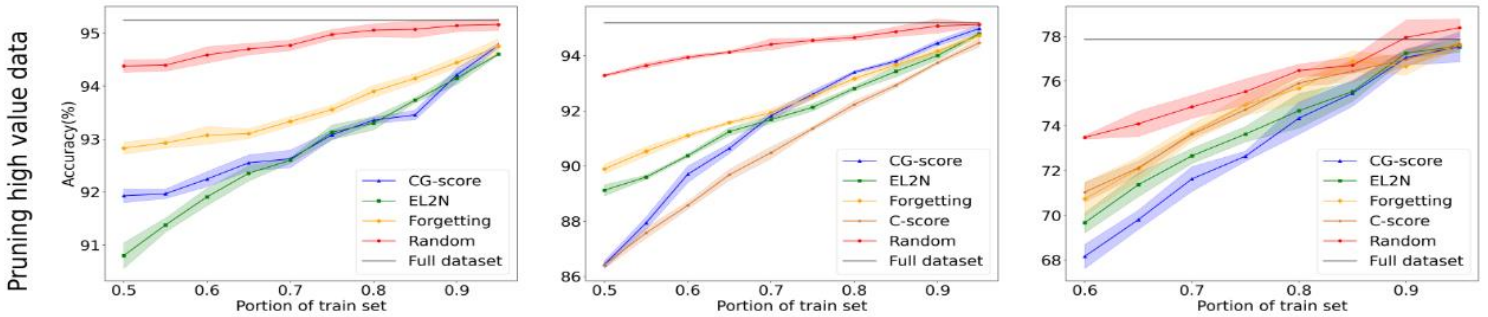
- 对比方法
  - Forgetting（2019）：对于一个特定的样本来说，他在t时刻被模型正确分类，而在t时刻之后，比如t+1时这个样本又被错误分类，则这个样本的遗忘事件发生次数+1
  - EL2N（2021）：这个样本优化模型参数之后，其他样本计算得到的Loss减少的量
  - C_score（2021）：通过量化标签数据的“结构性”，衡量了标签数据是否遵循某种特定的规律或关联模式，具有高度结构化的规律性的样本C_score值较高

CG_score 数据修剪实验

- 实验结果
  - 不需要对模型进行任何训练，可以与其他基线方法竞争性能
  - 首先删除高分样本时，测试精度最快下降，这意味着具有高CG分数的样本是控制模型拟合样本数据的重要组成部分



(a) Pruning low-scoring examples first. Better score maintains the test accuracy longer.

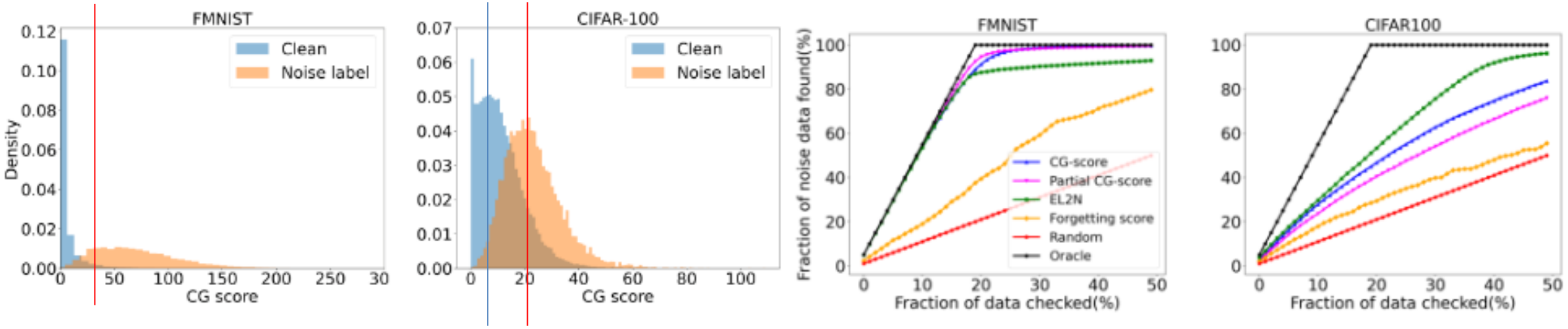


(b) Pruning high-scoring examples first. Better score makes the rapid performance drop.

为什么效果没有比基线好呢?

CG_score 噪声实验

- 实验结果
  - 具有损坏标签（橙色）的样本往往比具有干净标签（蓝色）的样本得分更高
  - 在更简单的数据集FMINIST（左）中，CG得分直方图可以在干净和噪音组之间更明显地分离
  - CG_score分数在简单数据集上也能够更好的检测噪声

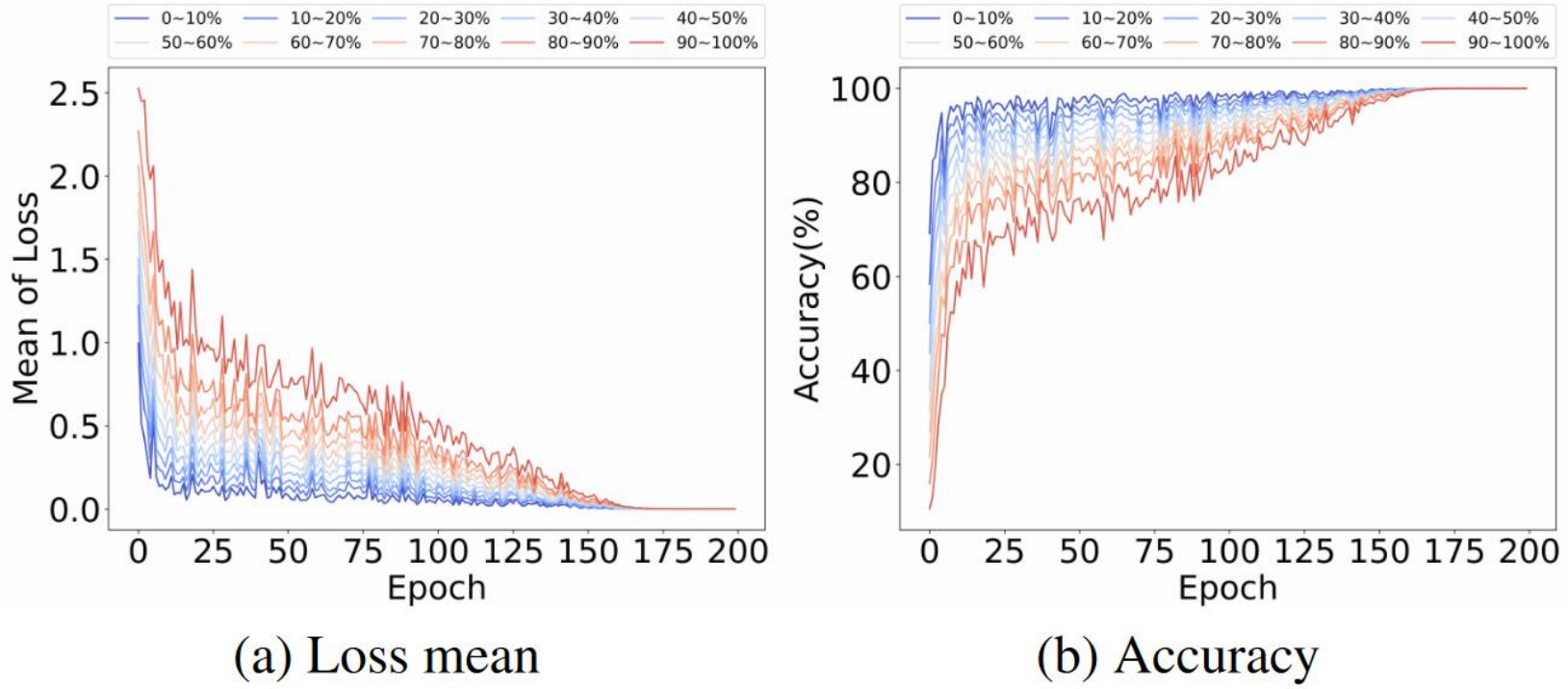


(a) Density of CG-score for clean vs. label-noise

(b) Fraction of detected label noise

CG_score 排序训练实验

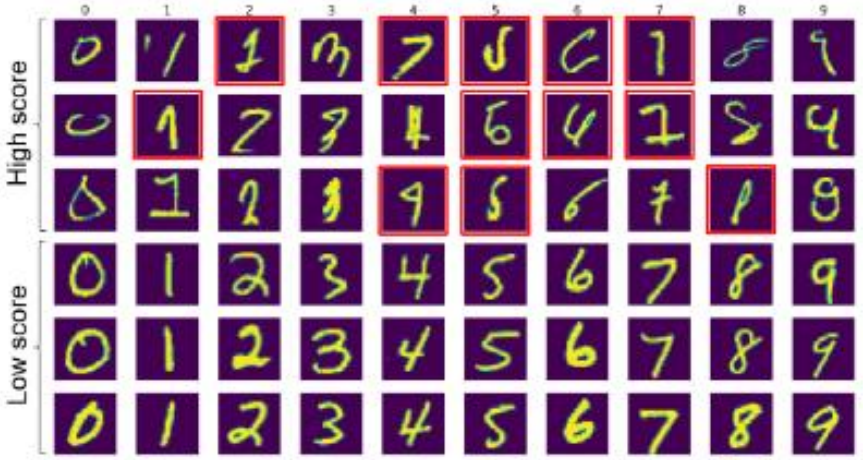
- 实验结果
  - 使用CG得分按升序排序数据实例，然后将数据分为10个相等大小的亚组
  - 随着训练的进展，我们测量10个亚组的损失和训练准确性
  - 可以观察到，低分组的平均损失小和准确率收敛快，而高分组的平均损失大和准确率收敛慢
  - 这表明CG分数与以模型学习速度衡量的示例“难度”高度相关



(a) Loss mean (b) Accuracy

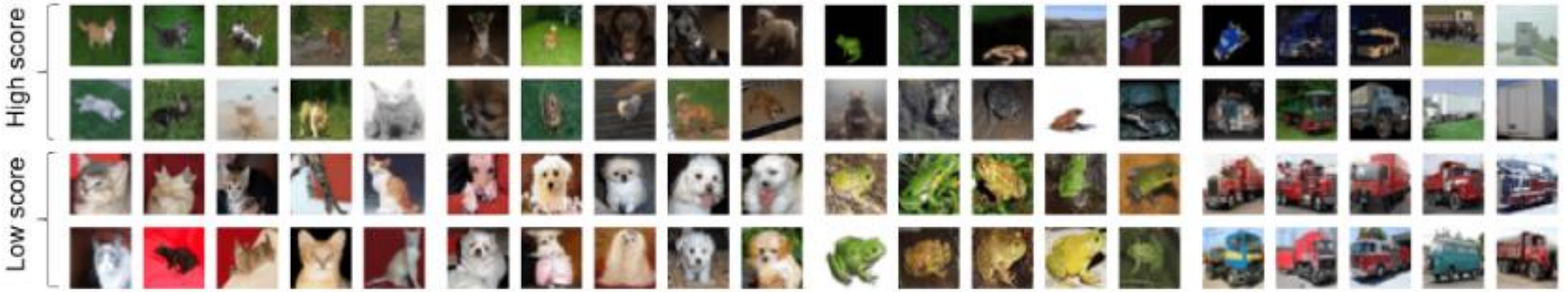
CG_score 可视化实验

- 实验结果
  - 最低CG得分的示例是代表每个类别的常规示例，它们看起来彼此相似
  - 而得分最高的示例则是不规则的，它们之间的外观不同



(a) Examples of MNIST dataset

(b) Examples of FMNIST dataset



(c) Examples of CIFAR-10 dataset (Cat, dog, frog, and truck)

特点总结与未来展望

特点总结与未来展望

- **特点总结**

优势	劣势
节省计算资源；提高效率	估算不够精确，准确性低
不依赖于特定的机器学习模型	无法全面评估数据的贡献

- **未来展望**
  - 多维度评估方法：应考虑数据的多维特性，如一致性、代表性等，以实现更全面、精准的评估
  - 统一规范的标准：目前数据样本重要性评估的标准和方法尚缺乏统一性，通过建立统一的标准框架，规定如何衡量和比较不同方法之间的相似性和差异
  - 提升可解释性与透明度

参考文献

[1] Nohyun K, Choi H, Chung H W. Data valuation without training of a model[C]. The Eleventh International Conference on Learning Representations. 2022.

[2] Arora S, Du S, Hu W, et al. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks[C]. International conference on machine learning. PMLR, 2019: 322-332.

[3] Ghorbani A, Zou J. Data shapley: Equitable valuation of data for machine learning[C]. International conference on machine learning. PMLR, 2019: 2242-2251.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！