基于可视分析的训练数据质量提升研究

杨维铠 [1)], 陈长建 [1)], 朱江宁 [1)], 李磊 [2)], 刘鹏 [2)], 刘世霞 [1)*]

[1)] (清华大学软件学院　北京　100084)
[2)] (中国航天科工集团第三研究院　北京　100083)
(shixia@tsinghua.edu.cn)

摘　要: 机器学习的成功依赖于高质量的训练数据. 但在实际应用中, 由于数据来源渠道多以及部分标注者水平不足, 训练数据质量很难得到保证. 为了解决这一问题, 可视分析技术通过深度结合机器学习和可视化技术, 将人融入到数据质量分析与提升回路中, 帮助提升训练数据质量, 从而提高模型性能. 本综述首先总结了训练数据质量问题的主要类型; 然后基于总结的问题类型, 对相关的可视分析工作进行分类与总结; 最后, 深入分析了基于可视分析的训练数据质量提升研究中所面临的机遇与挑战.

关键词: 机器学习; 数据质量; 可视分析; 可视化

中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2023.69

Visual Analytics Research for Improving Training Data Quality

Yang Weikai[1)], Chen Changjian[1)], Zhu Jiangning[1)], Li Lei[2)], Liu Peng[2)], and Liu Shixia[1)*]

[1)] (*School of Software, Tsinghua University, Beijing*　100084)
[2)] (*China Aerospace Science & Industry Corporation, Beijing*　100083)

Abstract: The success of machine learning relies on high-quality training data. However, it is difficult to ensure the quality of training data in practical applications due to the various sources of training data and the inexperience of some annotators. By tightly integrating machine learning and visualization, visual analytics techniques involve humans in the loop of data quality analysis and improvement, thereby enhancing the quality of training data and improving model performance. In this survey, we first summarize the main types of training data quality issues. Based on the identified problem types, we categorize and summarize relevant visual analytics approaches. Finally, we delve into the opportunities and challenges faced in research on training data quality improvement using visual analytics.

Key words: machine learning; data quality; visual analytics; visualization

随着大数据时代的到来, 机器学习变得愈发重要, 催生了众多应用, 如智慧医疗、智慧产线、自动驾驶、机器翻译、人脸识别等. 实践证明人工智能与机器学习成功的一个重要因素是高质量数据[1][2]. 例如, 机器学习在图像分类领域取得的快速进展, 很大程度上归功于高质量的标注数据

收稿日期: 2023-05-10; 修回日期: 20**-**-**. **基金项目**: 国家自然科学基金(U21A20469, 61936002); 国家重点研发计划(2020YFB2104100). **杨维铠**(1998—), 男, 博士研究生,主要研究方向为面向机器学习过程的可视分析; **陈长建**(1994—), 男, 博士研究生,主要研究方向为面向训练数据质量提升的可视分析; **朱江宁**(2000—), 男, 博士研究生,主要研究方向为大模型及其可视分析; **李磊**(1981—), 男, 博士, 研究员, 硕士生导师, 主要研究方向为飞行器设计; **刘鹏**(1975—), 男, 博士, 研究员, 硕士生导师, 主要研究方向为精确制导和智能感知; **刘世霞**(1974—), 女, 博士, 教授, 博士生导师, CCF 会员, 论文通讯作者, 主要研究方向为可解释机器学习、文本可视分析、文本挖掘.

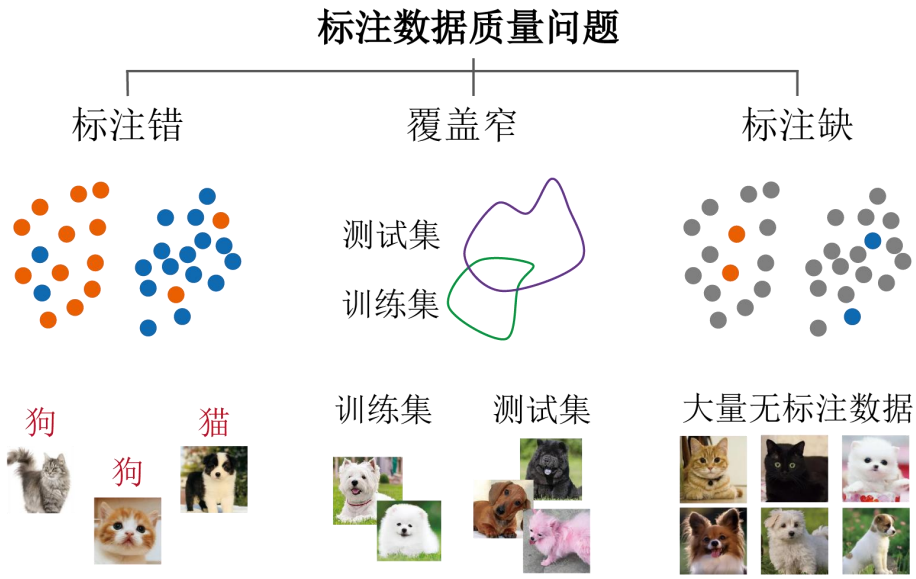


图 1. 三类主要的训练数据质量问题与典型示例.

ImageNet[3]. 在机器学习中，数据质量，尤其是训练数据的质量，决定了数据分析效果的上限[4]，而模型和算法的改进只是不断逼近这个上限.

然而在很多实际应用中，由于数据来源渠道多以及部分标注者水平不足，训练数据质量很难得到保证. 低质量训练数据会极大地降低模型的性能，进而给用户带来损失. 现有研究表明，与训练数据相关的质量问题主要包含以下三类[5]:

标注错，即数据的标注中存在错误，这导致模型在错误的监督信息下进行训练，最终造成误判. 据医学领域的调查显示，在美国每年至少有 4.4 万名患者因为数据标注错误引起的医疗事故而失去生命[6]. 而即使在科研领域里广泛应用的十个基准数据集中，也存在着 3.3%的错误标注[7].

覆盖窄，即训练数据分布未能有效覆盖测试数据分布. 例如，训练数据中仅包含白狗图像，但测试数据中包含各种颜色狗的图像. 这导致将白狗图像数据训练得到的模型用于预测其他颜色狗的类别时，性能出现急剧的下降. 著名工业信息咨询公司 The Data Warehousing Institute 的调查显示，在工业领域，覆盖窄的问题每年给工业公司带来了上亿美元的损失[8].

标注缺，即存在大量数据未被标注. 这是由于在很多实际应用中，数据产生速度远远大于数据标注速度，导致大量数据没有标注从而未能被有效利用，造成数据浪费. 据统计，在工业领域，未被有效标注和利用的数据高达 79%[9].

这三类训练数据质量问题及典型示例如图 1 所示.

近年来，国内外学者对可视分析和机器学习的交叉领域也进行了广泛的调研[10]-[19]. 其中 Yuan 等[18]系统地调研了 2010-2020 年间结合可视分析技术与机器学习的文章，并根据机器学习的流程从模型构建前、模型构建中、模型构建后三个阶段展开论述. 也有许多文章探索如何为机器学习任务设计更有效的可视分析系统[20]-[22]. 但这些工作没有从数据质量的角度进行详尽的分析. Liu 等[23]总结了多媒体数据、文本数据、轨迹数据和图数据中常用的数据清洗方法，并提出了一个通用的数据清洗流程. 但其并非针对训练数据，没有深入探讨数据质量问题对机器学习模型性能的影响及相应的解决方案. 为了深入探讨数据质量问题对机器学习模型性能的影响及相应的解决方案，本综述从训练数据质量问题的类型出发，对相关的可视分析工作进行系统性的分类与总结. 图 2 展示了本综述所涉及的文章范围和筛选流程，其采用了滚雪球的文献调研方式[24]. 本综述首先从 Liu 等[23]关于数据质量与可视分析的综述论文以及 Yuan 等[18]关于机器学习与可视分析的综述论文中所引用的论文出发，并人工检索了近三年发表在 VIS, EuroVis, PacificVis, CHI, TVCG, CGF, CG&A 等会议与期刊上包含“Visual Analytics”或“Visualization”且包含“Data quality”或“Training Data”关键词的论文，得到了 889 篇文章；随后人

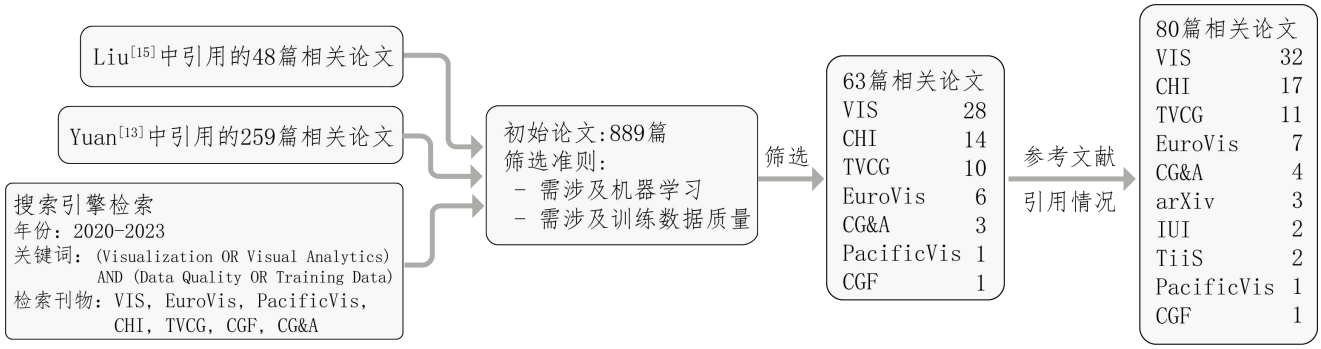
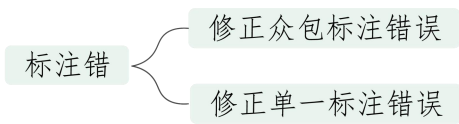


图 2. 本综述涉及的文章范围及筛选流程.

工排除了与机器学习和训练数据质量问题无关的论文，得到 63 篇相关论文. 再根据这些论文的参考文献和引用情况扩展了文章的范围，最终选定了 80 篇论文，并由三位博士生共同对每篇文章所解决的数据质量问题、数据质量提升方法以及使用的可视化形式进行了标注. 在此基础上，本综述针对标注错的问题，总结了标注错误修正方法；针对覆盖窄的问题，总结了数据集偏离纠正方法；针对标注缺的问题，总结了无标注数据质量提升方法. 图 3 汇总了不同方法所使用的可视化形式. 最后，本综述分析了基于可视分析的训练数据质量提升研究中所面临的机遇与挑战. 希望本综述能帮助机器学习领域和可视分析领域的研究人员增强对彼此工作的了解，进而将数据质量提升方法和可视分析技术进行更有效的结合. 本综述所涉及的相关论文也已经整理在 https://github.com/thu-vis/Visual-Analytics-Data-Quality 上，供相关研究人员参考.

1　标注错误修正方法



标注错误修正方法可根据每个样本的标注数量进一步分为众包标注错误修正方法和单一标注错误修正方法.

在众包标注数据中，每个样本由众包平台上的多个标注者进行标注. 这些众包标注往往包含较多的错误，需要使用众包标注错误修正方法来提升标注的质量. Park 等[25]提出了 C^{2}A 来可视化地展现众包标注者与其提供的众包标注的关系，进而帮助医学专家在医疗诊断视频中识别恶性肿瘤. 通过使用 C^{2}A，医学专家可以抛弃绝大多数不包含恶性肿瘤的影像片段，并将精力集中在包含恶性肿瘤的片段内. Park 等[26]进一步提出 CMed，帮助专家交互式地分析视频众包标注数据的准确性和标注者的标注能力. CMed 主要包括两个视图：标注视图和标注者视图 (图 4). 标注视图 (图 4(a)) 展示了标注视频帧占所有视频帧的比例、每个视频的标注区域等信息，帮助用户分析众包数据的准确性. 标注者视图 (图 4(b)) 展示了标注者的敏感度、在质量控制样本上的标注准确率以及标注的具体情况，从而帮助专家找到优秀的标注者，并剔除表现不好的标注者. 虽然 CMed 可以用来分析众包标注数据，但是其依赖于所有数据的真实标注，这在实际应用中往往难以获得. 在没有真实标注的情况下，为了修正众包标注错误，用户不仅需要检

数据质量问题	数据质量提升方法	散点图	网格布局	矩阵视图	表格	节点-链接图	图表
标注错（第1节）	修正众包标注错误（3）	[25][27]	[26]	[25][26][27]			[25][26]
	修正单一标注错误（8）	[28][29][30] [31][33][34]	[28][30][35]		[31][33][34]	[35]	[30][32][33]
覆盖窄（第2节）	检测偏离分布样本（30）	[37][40][44][45][46] [49][50][51][53][54] [55][57][58][61][62][65]	[36][38][39][40] [46][47][55][56] [57][59][60]	[47][51][56][57] [59][60][64][65]	[37][41][42] [43][45][48] [54][62][63]	[61][62]	[37][40][41][42][43][45] [48][50][51][52][53][54] [55][56][57][58][60][64][65]
	检测概念漂移样本（4）	[67][68]	[67][68]	[66][69]	[67]	[69]	[66][67][68][69]
标注缺（第3节）	辅助数据标注（30）	[72][73][75][77][78] [87][88][89][91][92] [93][96][97][98][99]	[70][71][72] [78][88]	[74][96][97]	[76][79][80][81] [84][90][92][94]	[88][95]	[71][73][74][78][81][83] [85][86][87][91][93][97]
	增强现有标注（5）	[100][101][102][103]	[102][103][104]	[100][101]	[100][102]	[104]	[103]

图 3. 数据质量提升可视分析方法的系统分类及使用的可视化技术汇总.

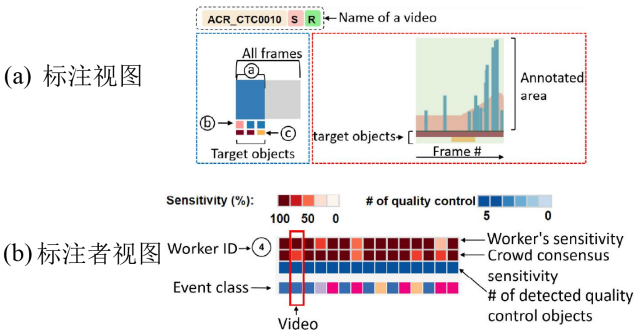


图 4. CMed[26]的两个主要视图：标注视图和标注者视图.图片来源于 Park 等[26]的工作. 已获得作者许可.

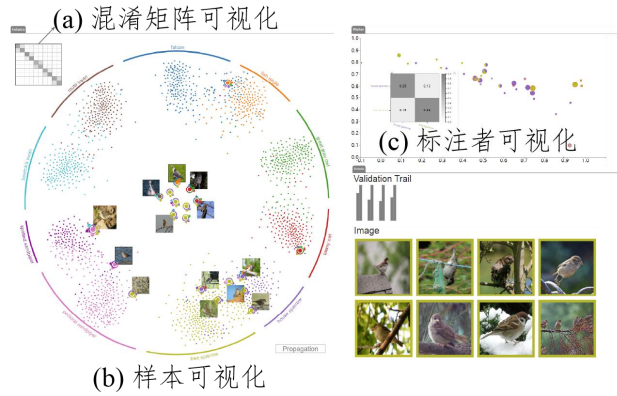


图 5. 众包数据标注错误修复可视分析工具 LabelInspect[27]. 图片来源于 Liu 等[27]的工作. 已获得作者许可.

查和确认样本标注，同时还要找到不可靠的标注者进行确认. 针对这一痛点问题, Liu 等[27]提出了一个可视分析方法 LabelInspect. 其首先使用众包学习模型对众包标注数据进行建模并预测样本的标注. 在此基础上，其使用互增强图模型对样本内部、标注者内部以及两者之间的影响进行建模，更好地推荐不确定样本标注和不可靠标注者给用户确认. LabelInspect 的可视化 (图 5) 主要包括三个模块：混淆矩阵可视化 (图 5(a))、样本可视化 (图 5(b)) 和标注者可视化 (图 5(c)). 混淆矩阵可视化用于展示不同类别之间的混淆程度，帮助用户选择易混淆的类别并对其进行进一步分析. 样本可视化使用有约束的 t-SNE 降维技术展示样本标注的不确定度以及它们的上下文. 标注者可视化以散点图形式展示每个标注者在所选定类别上的标注准确率以及无效标注评分，帮助用户发现和分析不可靠标注者. 用户的确认信息通过众包学习模型传播到其他未被确认的样本标注和标注者上，从而将确认信息的影响最大化，提高用户检查与修正的效率.

除了众包标注数据之外，许多数据集中每个样本仅有单一标注，如 ImageNet[3]. 由于这些单一标注的数据没有相关的众包信息 (例如标注者的标注情况)，上面的方法无法直接使用. 针对单一标注数据存在的标注错误问题, Xiang 等[28]开发了 DataDebugger (图 6)，通过利用用户确认的可信点集交互式地修正标注错误. 该方法将分层可视化与增量投影方法相结合，将训练数据投影到二维平面上，帮助用户有效探索大规模数据集并找到可信点进行确认. 这些可信点的信息随后被标注校正算法传播到整个数据集，从而修正数据中的标注错误. 许多工作还会根据模型预测结果和标注的不一致，推荐用户优先探索不一致较多的区域并进行修正[29]-[33]. 例如, Paiva 等[29]假设被错误分类的数据很可能包含错误标注. 基于这一假设，他们采用了多维投影增强的 Neighbor Joining Tree 来帮助用户探索错误分类的样本并纠正错误标注. 修正后，分类器使用修正后的标注进行改进，然后进行新一轮的修正. Bäuerle 等[31]则提出了三种指标来检测样本错误与标注错误. 然后以矩阵和散点图形式呈现这些错误，帮助用户理解和修正错误. Park 等[34]针对文本翻译任务中的平行语料，使用多个度量指标衡量其质量并使用散点图与平行坐标轴进行展示，用户可以从中筛选出存在潜在问题的语料进行检查与修正.

在上述标注错误修正方法中，用户最关心的是样本分布及标注情况. 因此散点图和网格布局被广泛应用于展现样本信息[27][35]. 此外，添加统计图表有助于用户从更宏观的角度了解当前数据集的标注错误情况，从而进行更有目的性的分析.

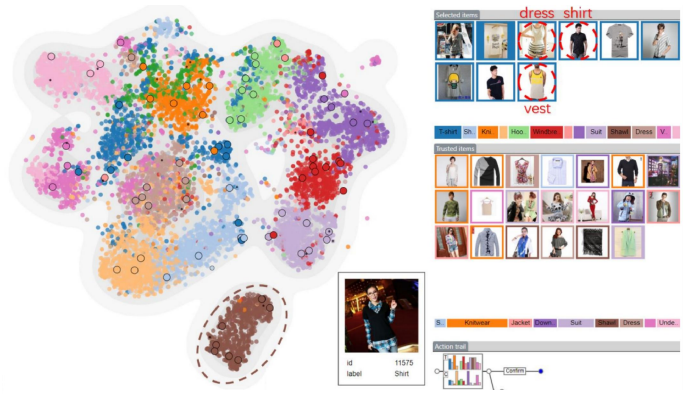
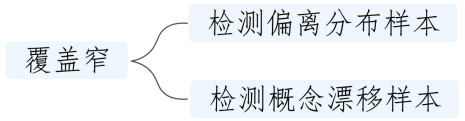


图 6. 基于可信点的标注错误修复可视分析工具 DataDebugger[28]. 图片来源于 Xiang 等[28]的工作. 已获得作者许可.

2　数据集偏离纠正方法



在机器学习中，有时会出现训练数据和测试数据中数据分布不一致的地方，即存在偏离分布样本[36]-[41]. 这些偏离分布样本会导致机器学习模型性能的下降. 为检测偏离分布样本，Chen 等[36]提出了一个可视分析方法 OoDAnalyzer (图 7). 其首先使用基于集成学习的偏离分布样本检测方法检测出候选偏离分布样本. 在检测到的候选偏离分布样本基础上，其使用网格布局来帮助用户探索与偏离分布样本相似的正常样本并查看样本内容，进而理解偏离分布样本出现的原因. Olson 等[38]则将与选择的数据相似的训练数据和测试数据分别展示在左右两侧，并根据偏离分布的程度从上往下排列，方便模型开发者更好地对比训练数据和测试数据分布的差异. Sharifi Noorian 等[39]开发了 Perspective，帮助模型开发者更高效地标注偏离样本(如出现在不常见场景中的物体)，为模型的训练提供更好的数据准备.

偏离分布样本也可以通过分析模型的预测结果进行检测. 模型开发者通过理解模型预测出现系统性偏差的原因，定位训练数据中存在的偏离分布样本，进而做出对应的修正[42]-[58]. 例如，Cao 等[43]展示了样本被模型逐层处理并最终得到预测结果的过程，帮助模型开发者分析正常样本与对抗样本预测结果出现差异的原因. Huang 等[56]开发了 ConceptExplainer，通过将图片分割为有语义的区域，更好地解释了模型的预测结果. 这一解释也帮助模型开发者更容易地发现数据集中的错误样本和偏离分布样本. 这一方式也被 Ahn 等[58]用于检测分类模型中的盲点并改进数据集. 有些工作还集成了训练数据构造方法，根据发现的偏离分布样本补充对应的训练数据. 例如，Gou 等[59]使用网格布局展现测试样本的分布，并使用热力图展现对应图片上目标检测的性能. 用户可通过分析性能较差的样本识别出训练样本集覆盖不足的情况，并利用基于语义的对抗学习方法有针对性地构造对抗样本，降低数据集的偏离程度. 分析性能不佳的原因并构造对抗样本的思路同样被成功应用于提升语义分割[60]和自然语言处理[61]-[63]等任务的训练数据质量. 除了构造新的样本，也可通过样本加权或重采样的方法来减轻数据集的偏离程度[64][65]. Zhang 等[64]开发的 SliceTeller 首先对数据进行自动划分，并推荐出性能不佳的数据子集. 模型开发者可尝试增大部分子集的权重以改善模型性能. 由于重新训练模型的代价十分高昂，该系统首先通过一个代理模型估计调整权重后模型性能的变化情况. 当模型开发者对这一变化感到满意后，模型才会在调整权重后的训练数据集上重新训练.

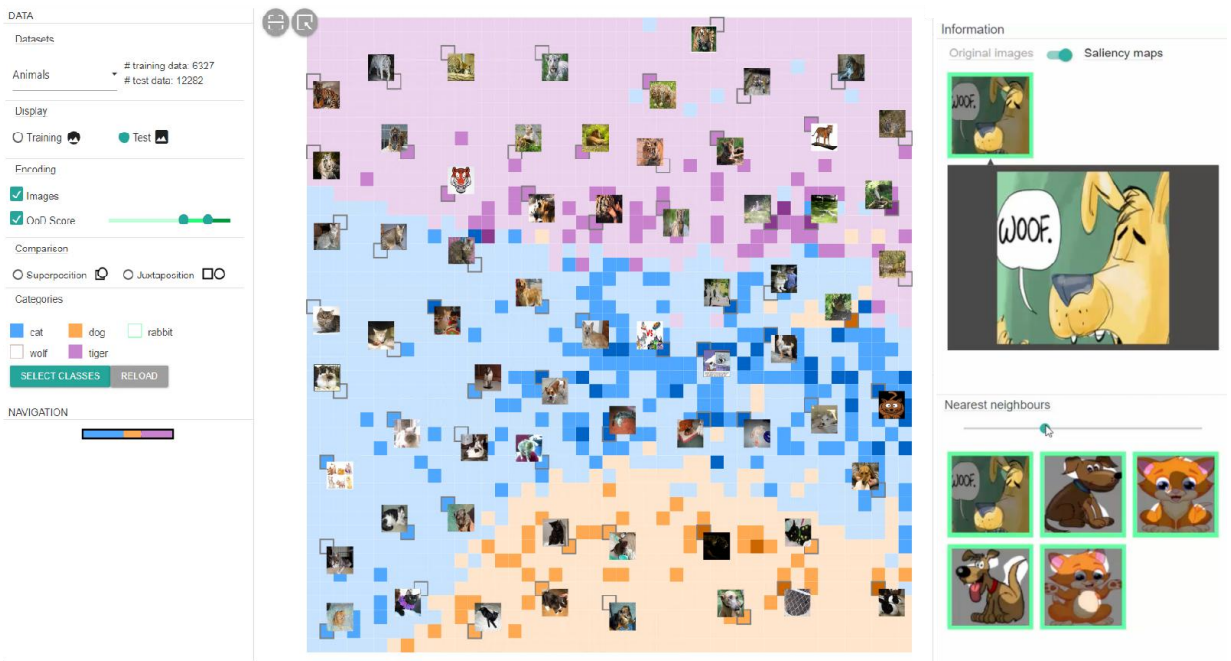


图 7. 偏离分布样本检测可视分析工具 OoDAnalyzer[36]. 图片来源于 Chen 等[36]的工作. 已获得作者许可.

在现实应用中，数据分布往往随着时间的推移不断发生变化，导致新到来的测试数据与训练数据分布不一致，这一现象称为概念漂移. 针对概念漂移，Wang 等[66]开发了 ConceptExplorer，根据预测模型的性能变化来监测概念漂移的发生，并通过多视图联动帮助用户对比分析多源时序数据中相关联的概念漂移. Yang 等[67]开发了 DriftVis (图 8). 该方法首先使用能量距离直接衡量当前测试数据分布与训练数据分布的概念漂移程度，并使用折线图展现概念漂移程度随时间的变化. 为了帮助用户更好地理解概念漂移发生的原因，其使用增量 t-SNE 技术将训练数据和不断到来的测试数据投影在同一个二维平面上. 用户进而可以快速定位概念漂移的时空特性并对其原因进行分析. 基于分析的结果，用户可以选择对应的数据加入到模型的训练过程中，从而完成对概念漂移的处理. Robertson 等[68]开发的 Angler 支持模型开发人员检查用户上传的数据并与模型的训练数据进行对比，从而发现之前训练数据覆盖不足的问题. 针对事件日志数据中发生的概念漂移，Yeshchenko 等[69]使用 DECLARE 模型建立了一系列约束，并据此对日志数据进行更细致的聚类分析与概念漂移检测.

在上述数据集偏离纠正方法中，用户最关心的是样本的分布以及训练/测试样本间的差异. 为此，将训练和测试样本投影到同一空间里进行分析是一个常用的手段[36][59]. 而为了突出训练/测试样本间的差异，通常会根据模型的输出[59]或样本间的距离[67]对其进行量化，并将其与样本一同展示[59][60]或使用单独图表进行可视化[66][67].

3 无标注数据质量提升方法

无标注数据质量提升的可视分析方法可分为两大类：一类辅助用户进行数据的推荐和标注，提高标注的效率与质量，这类方法通常不依赖于所使用的机器学习模型的工作机理；另一类则与半监督/无监督等模型紧密结合，通过可视分析方法增强现有的标注信息，最终提升模型性能.

早期关于数据推荐与标注的工作主要关注于提供高效的可视图表和交互方式，从而提高用户标注的效率. 其中最早的一份工作来自于 Moehrmann 等[70]. 他们开发了基于自组织映射的可视化. 在该可视化中，相似的图像被放置在一起，用户可以同时标注多个相似的图像，从而节省标注的时间和精力. 这个策略也被后续其他工作沿用和发展[71]-[75]. 例如，Kurzhals 等[71]将其用来标注眼动追踪数据；Halter 等[72]将其用来注释和分析电影中使用的颜色策略；Khayat 等[73]将其用来标注具有类似异常行为的垃圾社交信息机器人；Eirich 等[74]将其用来标注电动机的故障类型. Chang 等[75]证明了这一策略可以有效地提高非专业人员的标注准确率. 除了将相似的数据放置在

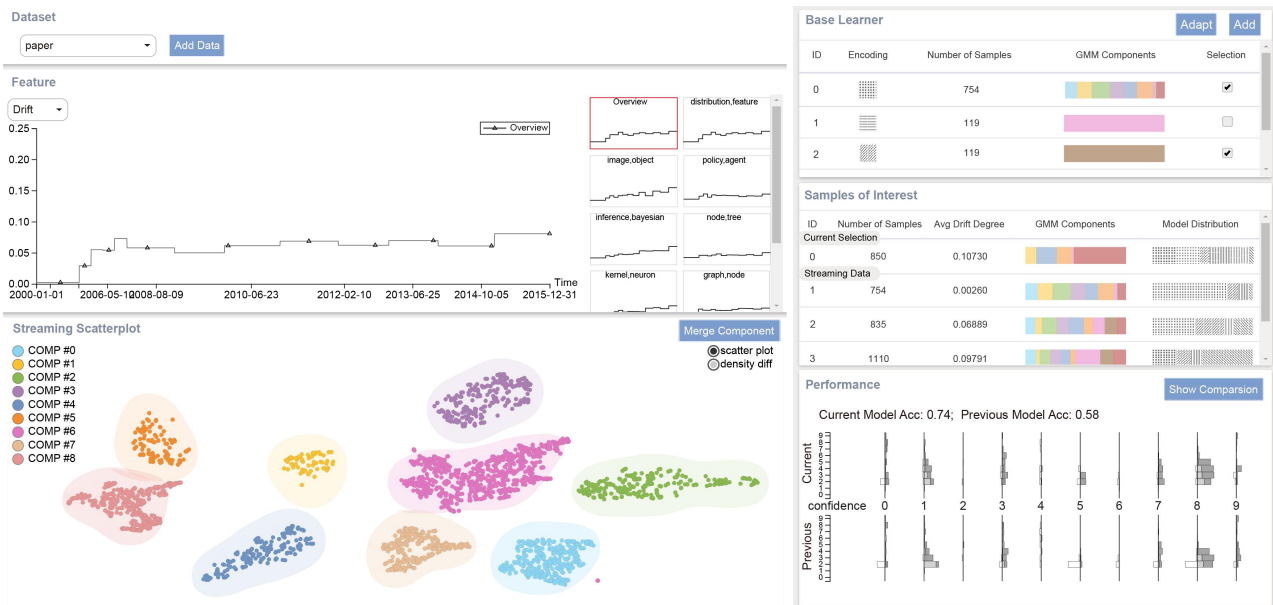
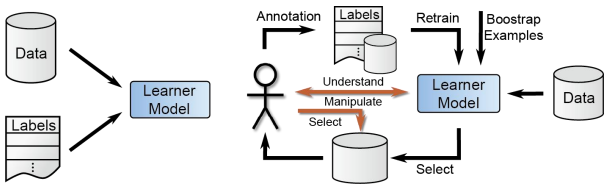


图 8. 概念漂移检测可视分析工具 DriftVis[67]. 图片来源于 Yang 等[67]的工作. 已获得作者许可.



(a) 传统模型训练流程 (b) 人在回路的模型训练流程

图 9. 传统模型训练流程和人在回路的模型训练流程的对比. 图片来源于 Höferlin 等[87]的工作. 已获得作者许可.

一起进行分析与标注以外，许多交互策略如过滤和排序等也被用来帮助用户查找感兴趣的样本进行标注. MediaTable[76]中利用过滤和排序来查找相似的视频片段. 其使用表格可视化来展现视频片段及其属性. 用户可以根据属性值过滤掉不相关的视频段，并对相关视频段进行排序与分析，从而允许用户同时标注属于同一类的多个视频段. Stein 等[77]提供了一个基于规则的过滤引擎来查找足球比赛视频中感兴趣的片段并标注. 用户可以通过自然语言 GUI 以交互方式指定规则并进行视频的筛选. 针对图像数据难以根据字段进行筛选、排序的问题，Hoque 等[78]提出了 Visual Concept Programming. 其首先将图片基于语义分割为若干片段，并分别为每个片段自动生成合适的语义标签. 用户可以根据语义标签设定标注规则，如“水”+“交通工具”=“船”. 用户可以检查并交互修改标注规则，最终完成对大规模数据的高效标注. 而针对众包数据标注场景，也有许多工作研究如何降低标注者的标注代价，并从众包标注中提取更高质量的类标[79]-[86]. 例如, Revolt[80]支持标注者标记存在歧义性的图片，并可以与其他标注者共享自己对于某个特定标注的解释. Gordon 等[86]将标注者的基本信息和标注历史展现出来，帮助模型开发者选择更无偏的标注者组合.

近期关于数据推荐与标注的工作则更关注于利用主动学习的方法给用户推荐出最需要标注的样本，并允许用户对推荐的结果进行分析与确认. 用户的确认信息会被传播到其他未标注的数据. 这类方法将机器和人紧密结合，发挥各自的优势. 这个方法被称为 Intra-active 标注方法 (图 9)，最早由 Höferlin 等[87]提出. 基于这一流程, Dennig 等[88]提出的可视分析系统 FDive 根据用户提供的标注信息，检测出对分类最有效的特征和距离函数，然后使用这些特征和距离函数来训练基于自组织映射的相关模型. 该模型可以在可视化上被用户直观地探索，并且可以通过获取更多标注来进一步

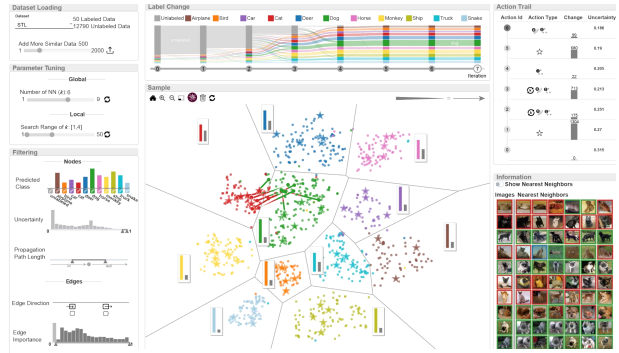


图 10. 改进图半监督学习中标注传播模式的可视分析工具 DataLinker[103]. 图片来源于 Chen 等[103]的工作. 已获得作者许可.

优化. 这一流程也被成功应用于多个应用场景，包括文本检索[89]，立场分类[90]，轨迹数据分类[91]，时序数据模式挖掘[93]，文本分类[93]，相关推特识别[94]等. Sperrle 等[95]开发了一种语言模型，其从文本中提取出需要进行标注的片段，并使用可视化帮助用户进行分析和标注. Yang 等[96]开发了 FSLDiagnotor 以帮助用户在小样本学习的场景下标注最关键的样本. 其使用散点图展现了当前样本的预测结果和预测置信度. Jia 等[97]将主动学习和零样本学习结合，通过让用户提供类别和属性之间对应关系的标注，交互式地构建一个高质量的类别-属性矩阵，从而提高零样本学习模型的性能. Bernard 等[98]的实验表明：以用户为中心的交互式标注优于以模型为中心的主动学习. Bernard 等[99]还对用户在标注过程中选择样本的策略进行了定量分析. 结果表明基于数据的策略 (例如集群、密集区域) 在标注早期效果很好，而基于模型的用户策略 (例如类别分离) 在后期表现更好.

随着弱监督学习的迅速发展，将可视分析与相关的机器学习模型紧密结合成为提升无标注数据质量的一个解决思路. 许多工作将迁移学习和可视分析紧密结合[100]-[102]，帮助模型开发者理解迁移学习的过程，并选择更好的训练数据进行迁移学习，进而改善模型性能. 半监督学习则通过将标注传播到无标注样本以改善模型性能. 针对传播过程中可能出现错误的问题, Chen 等[103]开发了 DataLinker 帮助用户探索、理解并改进图半监督学习模型中的标注传播过程. 其使用一个河流隐喻显示了标注传播模式的概览，并结合散点图、结点连接图和条形图展示样本的空间分布 (图 10). 这两个可视化协同工作，帮助用户找出有问题的样本以供用户进一步分析和修改. 相关的修改被图

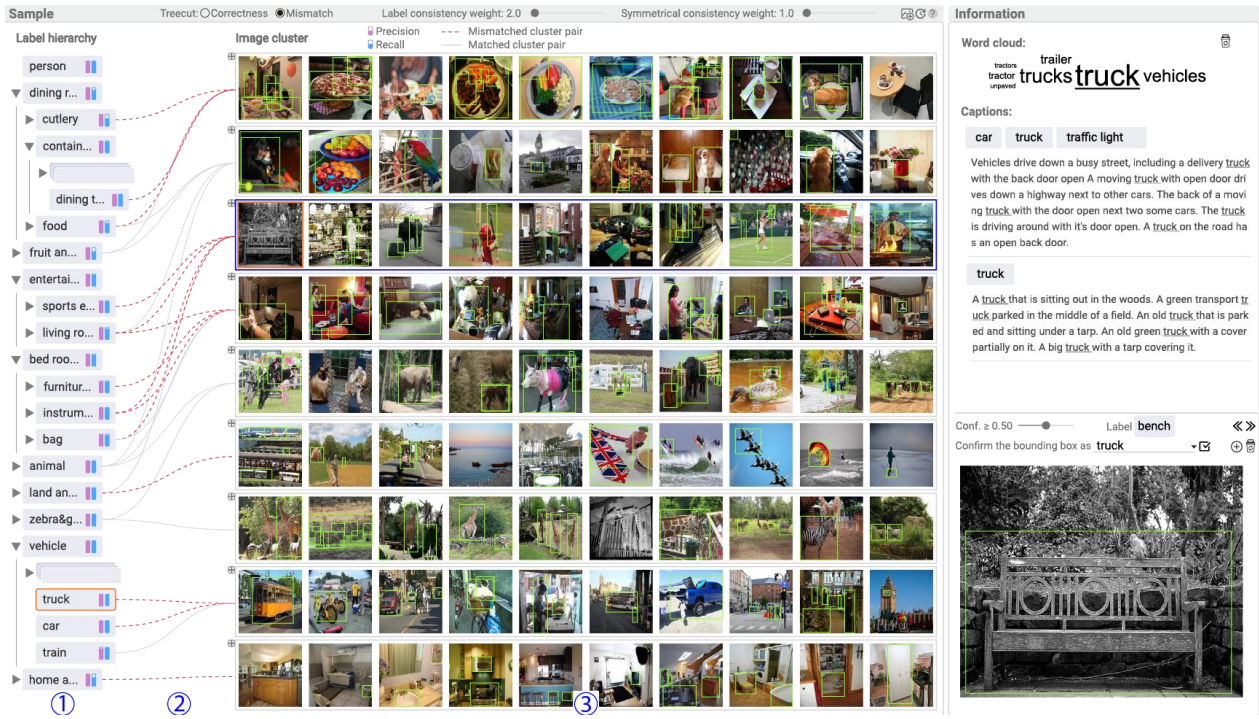


图 11. 利用图像说明改进标注的可视分析工具 MutualDetector[104]. 图片来源于 Chen 等[104]的工作. 已获得作者许可.

半监督模型进一步利用, 从而提升无标注数据质量并提高模型性能. 除了通过传播已有的数据标注信息进行标注数据质量的提升, 也有一类方法利用其他模态的信息来提升标注质量. 例如, 尽管互联网上许多图片没有精确的标注, 但其对应的图像说明通常包含标注的信息. Chen 等[104]提出了 MutualDetector (图 11), 综合利用图像目标检测的结果和图像说明中的类别信息进行迭代式的改进. 用户可以通过一个基于节点-连接的集合可视化, 直观地探索目标检测的结果和提取的标注信息的共现关系, 定位当前模型性能不佳的瓶颈并进行调整, 使得两者的信息能更好地相互增强, 提升数据质量.

在上述无标注数据质量提升方法中, 用户最关心的是如何提高标注的效率和质量. 通过使用保持样本相似性的可视化方法[70][76]可以在标注过程中提供丰富的上下文信息, 进而提高用户的标注效率并提升标注质量. 而根据用户的反馈更智能地进行样本选择和推荐[87][96]也是进一步提高标注效率的常用手段.

4 研究机遇与挑战

综上所述, 可视分析技术已经在训练数据质量提升中成功应用, 并体现了突出的技术优势. 针对“标注错”问题, 现有工作集中在检测并修正分类问题中的数据标注错误. 针对“覆盖窄”问题, 现有工作给出了对偏离分布样本以及概念漂移样本的检测与修正方法. 针对“标注缺”问题, 现有工作通过批量标注与标注传播等方法实现了高效的数据标注. 从这些工作中可以看到以数据为中心的可视分析方法越来越受到重视. 其中大多数工作都会采用散点图和网格布局等技术直接展现数据的分布及其内容, 并通过后端的模型与算法分析数据质量问题并给出可视的推荐和引导, 降低用户分析数据的门槛. 而用户的反馈也将被后端的模型和算法充分利用, 提高用户分析的效率. 但目前该领域仍有一些具有挑战性的问题待解决, 主要包含以下几个方面:

复杂任务中的数据质量问题. 目前大多数数据质量提升工作都是针对图像分类任务中图片标注存在的质量问题, 如标注错误或缺失. 而在机器学习领域中, 许多其他任务也依赖于高质量的训练数据, 这些数据的标注更加复杂. 例如图像分割任务不仅需要类别信息, 还需要像素级别的标注信息, 机器翻译任务需要双语句子对等. 这些任务中训练数据的理解、分析与修正都比图像分类任务更加复杂. 如何有效地结合可视分析技术, 为不同任务所需的训练数据提供有效的数据质量提升方法值得更深入的研究.

大语言模型中的数据质量问题. 由于其极佳的性能和易于使用的特性，以 GPT-4 为代表的大语言模型得到广泛关注并迅速发展[105][106]. 基于小样本的上下文学习是目前将大语言模型应用于下游任务有效方法[107]. 它将少量的训练样本转化为提示 (prompt)，并和下游任务一起送入模型进行推理. 现有研究表明这些训练样本的标注[108][109]、代表性[110]和顺序[111]都会影响下游任务的性能，但目前仍缺乏有效的手段理解这些样本对性能的影响并进行有针对性的改进. 因此，如何结合可视分析技术，揭示上下文学习的工作机理，指导模型开发者更有效地选择高质量训练样本作为上下文是一个未来的研究方向.

多模态数据质量问题. 目前，数据质量提升方法主要关注单模态数据，如表格数据、图像数据等. 在实际应用中，数据往往以多模态形式存在. 例如，视频数据中既包含图像数据，也包含音频数据. 医疗记录既包含图像数据 (如 CT、MRI)，也包含文本数据 (如病历). 多模态之间可以相互增强和融合，从而更好地提升训练数据质量. 然而，现有的多模态融合方法大多基于数据质量高的假设，难以适用于数据质量低的场景. 为此，如何在数据质量低的情况下对多模态数据进行建模和融合，并通过可视分析的手段提升数据质量是一个未来的研究方向.

流数据质量问题. 后续工作中的另一个研究重点是流数据中的质量问题. 在实际应用中，数据总是不断地产生，具有数据量大、产生速度快、数据质量问题不断变化等特点，因此需要实时处理数据质量问题. 然而，现有流数据处理方法不能很好地处理高通量大规模的数据，也没有涵盖多种数据质量问题. 为此，需要研究更高效的增量式数据建模方法[67][112]，从而可以对新数据中的质量问题快速准确地检测，并设计对应的增量式的可视化方法，实时展示新到来的数据并揭示其中的数据质量问题，帮助用户进行分析与处理.

多种数据质量问题交织. 目前，大部分标注数据质量提升方法都假设数据中仅存在一种标注数据质量问题. 然而，在很多实际应用中，一个标注数据集往往存在多种质量问题. 例如，医疗数据中往往同时存在覆盖窄和标注缺的质量问题. 多种标注数据质量问题交织在一起，给检测和分析带来了困难. 检测多种标注数据质量问题需要更加复杂的模型建模以及更加灵活的知识融合框架. 分析这些问题则需要更加合适的设计和布局来有效展示数据与存在的多种质量问题. 如何解决以上问题仍然需要进一步研究.

5 结 语

本综述首先总结了三类常见的训练数据质量问题，并根据这三类问题对训练数据质量提升相关的可视分析工作系统性地进行分类与总结，深入分析了该研究领域所面临的机遇与挑战. 希望本综述能够帮助读者对基于可视分析的训练数据质量提升研究建立更全面而深入的认识，并启发相关工作的进一步研究.

参考文献(References):

[1] Tian T, Zhu J. Max-Margin Majority Voting for Learning From Crowds[C] //Proceedings of the Advances in Neural Information Processing Systems, 2015: 1621-1629.

[2] Liu M, Jiang L, Liu J, et al. Improving Learning-from-crowds Through Expert Validation[C] //Proceedings of the International Joint Conference on Artificial Intelligence, 2017: 2329-2336.

[3] Deng J, Dong W, Socher R, et al. ImageNet: A Large-scale Hierarchical Image Database[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.

[4] Crotes C, Jackel L D, Chiang W P. Limits on Learning Machine Accuracy Imposed by Data Quality[C] //Proceedings of the Advances in Neural Information Processing Systems, 1995: 239-246.

[5] 李建中，王宏志，高宏. 大数据可用性的研究进展[J]. 软件学报, 2016, 27(7): 1605-1625.

[6] Donaldson M S, Corrigan J M, Kohn L T, et al. To Err is Human: Building a Safer Health System[M]. National Academy Press, 2000.

[7] Northcutt C G, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks[C] //Proceedings of Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.

[8] Eckerson W. Data Warehousing Special Report: Data Quality and the Bottom Line[J]. Applications Development Trends, 2002, 1(1): 1-9.

[9] Woodie A. Data management: Still a Major Obstacle to AI Success[EB/OL]. (2019-05-22)[2022-02-25]. https://www.datanami.com/2019/05/22/data-management-still-a-major-obstacle-to-ai-success/.

[10] Liu S, Wang X, Liu M, et al. Towards Better Analysis of Machine Learning Models: A Visual Analytics Perspective[J]. Visual Informatics, 2017, 1(1): 48–56

[11] Hohman F, Kahng M, Pienta R, et al. Visual Analytics in Deep

Learning: An Interrogative Survey for the Next Frontiers[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(8): 2674–2693

[12] Sacha D, Kraus M, Keim D A, et al. VIS4ML: An Ontology for Visual Analytics Assisted Machine Learning[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 385–395.

[13] Jiang L, Liu S, Chen C. Recent Research Advances on Interactive Machine Learning[J]. Journal of Visualization, 2019, 22(2): 401-417.

[14] Chatzimparmpas A, Martins R M, Jusufi I, et al. The state of the art in enhancing trust in machine learning models with the use of visualizations[C] //Computer Graphics Forum, 2020, 39(3): 713-756.

[15] La Rosa B, Blasilli G, Bourqui R, et al. State of the art of visual analytics for explainable deep learning[C] //Computer Graphics Forum, 2023, 42(1): 319-355.

[16] Sambasivan N, Kapania S, Highfill H, et al. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI[C] //proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021: 1-15.

[17] Gil Y, Honaker J, Gupta S, et al. Towards human-guided machine learning[C] //Proceedings of the International Conference on Intelligent User Interfaces. 2019: 614-624.

[18] Yuan J, Chen C, Yang W, et al. A Survey of Visual Analytics Techniques for Machine Learning[J]. Computational Visual Media, 2021, 7(1): 3-36.

[19] 夏佳志, 李杰, 陈思明, 等. 可视化与人工智能交叉研究综述[J]. 中国科学: 信息科学, 2021.

[20] Wang Q, Huang K, Chandak P, et al. Extending the nested model for user-centric XAI: a design study on GNN-based drug repurposing[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(1): 1266-1276.

[21] Zhang Y, Wang Y, Zhang H, et al. Onelabeler: A flexible system for building data labeling tools[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2022: 1-22.

[22] Bäuerle A, Cabrera Á A, Hohman F, et al. Symphony: Composing interactive interfaces for machine learning[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2022: 1-14.

[23] Liu S, Andrienko G, Wu Y, et al. Steering Data Quality with Visual Analytics: The Complexity Challenge[J]. Visual Informatics, 2018, 2(4): 191-197.

[24] Wohlin C. Guidelines for snowballing in systematic literature studies and a replication in software engineering[C]//Proceedings of the international conference on evaluation and assessment in software engineering. 2014: 1-10.

[25] Park J H, Nadeem S, Mirhosseini S, et al. C^2A: Crowd Consensus Analytics for Virtual Colonoscopy[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology, 2016: 21-30.

[26] Park J H, Nadeem S, Boorboor S, et al. CMed: Crowd Analytics for Medical Imaging Data[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(6): 2869-2880.

[27] Liu S, Chen C, Lu Y, et al. An Interactive Method to Improve Crowdsourced Annotations[J]. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 235-245.

[28] Xiang S, Ye X, Xia J, et al. Interactive Correction of Mislabeled Training Data[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology. 2019: 57-68.

[29] Paiva J G S, Schwartz W R, Pedrini H, et al. An Approach to Supporting Incremental Visual Data Classification[J]. IEEE Transactions on Visualization and Computer Graphics, 2015, 21(1): 4-17.

[30] Garcia Caballero H S, Westenberg M A, Gebre B, et al. V-Awake: A Visual Analytics Approach for Correcting Sleep Predictions from Deep Learning Models[C] //Computer Graphics Forum, 2019, 38(3): 1-12.

[31] Bäuerle A, Neumann H, Ropinski T. Classifier-Guided Visual Correction of Noisy Labels for Image Classification Tasks[J]. Computer Graphics Forum, 2020, 39(3): 195-205.

[32] Nourani M, Roy C, Honeycutt D R, et al. DETOXER: A Visual Debugging Tool With Multiscope Explanations for Temporal Multilabel Classification[J]. IEEE Computer Graphics and Applications, 2022, 42(6): 37-46.

[33] Zhang X, Xuan X, Dima A, et al. LabelVizier: Interactive Validation and Relabeling for Technical Text Annotations[J]. arXiv preprint arXiv:2303.17820, 2023.

[34] Park S, Lee S, Kim Y, et al. VANT: A Visual Analytics System for Refining Parallel Corpora in Neural Machine Translation[C] //Proceedings of the Pacific Visualization Symposium. IEEE, 2022: 181-185.

[35] Liu M, Shi J, Li Z, et al. Towards better analysis of deep convolutional neural networks[J]. IEEE transactions on visualization and computer graphics, 2016, 23(1): 91-100.

[36] Chen C, Yuan J, Lu Y, et al. OoDAnalyzer: Interactive Analysis of Out-of-Distribution Samples[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(7): 3335-3349.

[37] Song D, Wang Z, Huang Y, et al. DeepLens: Interactive Out-of-distribution Data Detection in NLP Models[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2023: 1-17.

[38] Olson M L, Nguyen T V, Dixit G, et al. Contrastive identification of covariate shift in image data[C] //IEEE Visualization Conference (VIS). IEEE, 2021: 36-40.

[39] Sharifi Noorian S, Qiu S, Sayin B, et al. Perspective: Leveraging Human Understanding for Identifying and Characterizing Image Atypicality[C] //Proceedings of the International Conference on Intelligent User Interfaces. 2023: 650-663.

[40] Wang X, Chen W, Xia J, et al. HetVis: A Visual Analysis Approach for Identifying Data Heterogeneity in Horizontal Federated Learning[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(1): 310-319.

[41] Narechania A, Du F, Sinha A R, et al. DataPilot: Utilizing Quality and Usage Information for Subset Selection during Visual Data Preparation[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2023: 1-18.

[42] Liu M, Liu S, Su H, et al. Analyzing the noise robustness of deep neural networks[C] //IEEE Conference on Visual Analytics Science and Technology. 2018: 60-71.

[43] Cao K, Liu M, Su H, et al. Analyzing the noise robustness of deep neural networks[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(7): 3289-3304.

[44] Rathore A, Dev S, Phillips J M, et al. VERB: Visualizing and Interpreting Bias Mitigation Techniques for Word Representations[J]. arXiv preprint arXiv:2104.02797, 2021.

[45] Bäuerle A, Turker A G, Burke K, et al. Visual Identification of Problematic Bias in Large Label Spaces[J]. arXiv preprint arXiv:2201.06386, 2022.

[46] Wang Q, Xu Z, Chen Z, et al. Visual analysis of discrimination in machine learning[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1470-1480.

[47] Xie T, Ma Y, Kang J, et al. FairRankVis: A visual analytics framework for exploring algorithmic fairness in graph mining models[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 368-377.

[48] Cheng F, Ming Y, Qu H. DECE: Decision explorer with counterfactual explanations for machine learning models[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1438-1447.

[49] Munechika D, Wang Z J, Reidy J, et al. Visual Auditor: Interactive Visualization for Detection and Summarization of Model Biases[C] //IEEE Visualization and Visual Analytics (VIS Short Papers). IEEE, 2022: 45-49.

[50] Kwon B C, Kartoun U, Khurshid S, et al. RMExplorer: A Visual Analytics Approach to Explore the Performance and the Fairness of Disease Risk Models on Population Subgroups[C] //IEEE Visualization and Visual Analytics (VIS Short Papers). IEEE, 2022: 50-54.

[51] Arunkumar A, Sharma S, Agrawal R, et al. LINGO: Visually Debiasing Natural Language Instructions to Support Task Diversity[J] Computer Graphics Forum, 2023. To be published.

[52] Ghai B, Mueller K. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(1): 473-482.

[53] Arendt D L, Nur N, Huang Z, et al. Parallel embeddings: a visualization technique for contrasting learned representations[C] //Proceedings of the International Conference on Intelligent User Interfaces. 2020: 259-274.

[54] Ma Y, Xie T, Li J, et al. Explaining vulnerabilities to adversarial machine learning through visual analytics[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 1075-1085.

[55] Li Y, Wang J, Fujiwara T, et al. Visual Analytics of Neuron Vulnerability to Adversarial Attacks on Convolutional Neural Networks[J]. ACM Transactions on Interactive Intelligent Systems, 2023.

[56] Huang J, Mishra A, Kwon B C, et al. ConceptExplainer: Interactive Explanation for Deep Neural Networks from a Concept Perspective[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(1): 831-841.

[57] Kwon B C, Lee J, Chung C, et al. DASH: Visual Analytics for Debiasing Image Classification via User-Driven Synthetic Data Augmentation[C]// Eurographics Conference on Visualization (Short Papers) 2022: 91-95

[58] Ahn Y, Lin Y R, Xu P, et al. ESCAPE: Countering Systematic Errors from Machine's Blind Spots via Interactive Visual Analysis[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2023: 1-16.

[59] Gou L, Zou L, Li N, et al. VATLD: A Visual Analytics System to Assess, Understand and Improve Traffic Light Detection[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 261-271.

[60] He W, Zou L, Shekar A K, et al. Where can we help? a visual analytics approach to diagnosing and improving semantic segmentation of movable objects[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 1040-1050.

[61] Li Z, Wang X, Yang W, et al. A unified understanding of deep nlp models for text classification[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(12): 4980-4994.

[62] Jin Z, Wang X, Cheng F, et al. ShortcutLens: A Visual Analytics Approach for Exploring Shortcuts in Natural Language Understanding Dataset[J]. IEEE Transactions on Visualization and Computer Graphics, 2023. To be published.

[63] Wu S, Shen H, Weld D S, et al. ScatterShot: Interactive In-context Example Curation for Text Transformation[C] //Proceedings of the International Conference on Intelligent User Interfaces. 2023: 353-367.

[64] Zhang X, Ono J P, Song H, et al. SliceTeller: A Data Slice-Driven Approach for Machine Learning Model Validation[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 29(1): 842-852.

[65] Chatzimparmpas A, Paulovich F V, Kerren A. HardVis: Visual Analytics to Handle Instance Hardness Using Undersampling and Oversampling Techniques[J] Computer Graphics Forum, 2023, 42(1): 135-154.

[66] Wang X, Chen W, Xia J, et al. ConceptExplorer: Visual Analysis of Concept Drifts in Multi-source Time-series Data[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology. 2020: 1-11.

[67] Yang W, Li Z, Liu M, et al. Diagnosing Concept Drift with Visual Analytics[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology. 2020: 12-23.

[68] Robertson S, Wang Z J, Moritz D, et al. Angler: Helping Machine Translation Practitioners Prioritize Model Improvements[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2023: 1-20.

[69] Yeshchenko A, Di Ciccio C, Mendling J, et al. Visual Drift Detection for Event Sequence Data of Business Processes[J]. IEEE Transactions on Visualization and Computer Graphics. 2022. 28(8): 3050-3068.

[70] Moehrmann J, Bernstein S, Schlegel T, et al. Improving the Usability of Hierarchical Representations for Interactively Labeling Large Image Data Sets[C] //Proceedings of the International Conference on Human-Computer Interaction. 2011: 618-627.

[71] Kurzhals K, Hlawatsch M, Seeger C, et al. Visual analytics for mobile eye tracking[J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1): 301-310.

[72] Halter G, Ballester-Ripoll R, Flueckiger B, et al. VIAN: A Visual Annotation Tool for Film Analysis[J]. Computer Graphics Forum, 2019, 38(3): 119-129.

[73] Khayat M, Karimzadeh M, Zhao J, et al. VASSL: A Visual Analytics Toolkit for Social Spambot Labeling[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 874-883.

[74] Eirich J, Bonart J, Jäckle D, et al. IRVINE: A Design Study on Analyzing Correlation Patterns of Electrical Engines[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 11-21.

[75] Chang C M, Lee C H, Igarashi T. Spatial labeling: leveraging spatial layout for improving label quality in non-expert image annotation[C]//Proceedings of the CHI Conference on Human Factors in Computing Systems. 2021: 1-12.

[76] Rooij O, van Wijk J, Worring M. MediaTable: Interactive Categorization of Multimedia Collections[J]. IEEE Computer Graphics and Applications, 2010, 30(5): 42-51.

[77] Stein M, Janetzko H, Breitkreutz T, et al. Director's cut: Analysis and Annotation of Soccer Matches[J]. IEEE Computer Graphics and Applications, 2016, 36(5): 50-60.

[78] Hoque M N, He W, Shekar A K, et al. Visual Concept Programming: A Visual Analytics Approach to Injecting Human Intelligence at Scale[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(1): 74-83.

[79] Mitra T, Hutto C J, Gilbert E. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk[C] //Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015: 1345-1354.

[80] Chang J C, Amershi S, Kamar E. Revolt: Collaborative crowdsourcing for labeling machine learning datasets[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2017: 2334-2346.

[81] Barbosa N M, Chen M. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2019: 1-12.

[82] Cartwright M, Dove G, Méndez Méndez A E, et al. Crowdsourcing multi-label audio annotation tasks with citizen scientists[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2019: 1-11.

[83] Saha M, Saugstad M, Maddali H T, et al. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2019: 1-14.

[84] Méndez Méndez A E, Cartwright M, Bello J P, et al. Eliciting Confidence for Improving Crowdsourced Audio Annotations[J]. Proceedings of the ACM on Human-Computer Interaction, 2022, 6(CSCW1): 1-25.

[85] Kong N, Hearst M A, Agrawala M. Extracting references between text and charts via crowdsourcing[C] //Proceedings of the CHI conference on Human Factors in Computing Systems. 2014: 31-40.

[86] Gordon M L, Lam M S, Park J S, et al. Jury learning: Integrating dissenting voices into machine learning models[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2022: 1-19.

[87] Höferlin B, Netzel Rch, Höferlin M, et al. Interactive Learning of Ad-hoc Classifiers for Video Visual Analytics[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology, 2012: 23-32.

[88] Dennig F L, Polk T, Lin Z, et al. FDive: Learning Relevance Models using Pattern-based Similarity Measures[C] //Proceedings of the IEEE Conference on Visual Analytics Science and Technology. 2019: 69-80.

[89] Heimerl F, Koch S, Bosch H, et al. Visual Classifier Training for Text Document Retrieval[J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(12): 2839-2848.

[90] Kucher K, Paradis C, Sahlgren M, et al. Active learning and visual analytics for stance classification with ALVA[J]. ACM Transactions on Interactive Intelligent Systems, 2017, 7(3): 1-31.

[91] Júnior A S, Renso C, Matwin S. Analytic: An Active Learning System for Trajectory Classification[J]. IEEE Computer Graphics and Applications, 2017, 37(5): 28-39.

[92] Choi M, Park C, Yang S, et al. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks[C] //Proceedings of the CHI conference on human factors in computing systems. 2019: 1-12.

[93] Lekschas F, Peterson B, Haehn D, et al. Peax: Interactive Visual Pattern Search in Sequential Data Using Unsupervised Deep Representation Learning[J]. Computer Graphics Forum. 2020, 39(3): 167-179.

[94] Snyder L S, Lin Y S, Karimzadeh M, et al. Interactive Learning for Identifying Relevant Tweets to Support Real-time Situational Awareness[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1): 558-568.

[95] F. Sperrle, R. Sevastjanova, R. Kehlbeck, et al. VIANA: Visual Interactive Annotation of Argumentation[C] //Proceedings of the Conference on Visual Analytics Science and Technology, 2019, 11–22.

[96] Yang W, Ye X, Zhang X, et al. Diagnosing ensemble few-shot classifiers[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(9): 3292-3306.

[97] Jia S, Li Z, Chen N, et al. Towards Visual Explainable Active learning for Zero-shot Classification[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(1): 791-801.

[98] Bernard J, Hutter M, Zeppelzauer M, et al. Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study[J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 24(1): 298-308.

[99] Bernard J, Zeppelzauer M, Lehmann M, et al. Towards User-Centered Active Learning Algorithms[J]. Computer Graphics Forum. 2018, 37(3): 121-132.

[100] Ma Y, Fan A, He J, et al. A visual analytics framework for explaining and diagnosing transfer learning processes[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(2): 1385-1395.

[101] Mishra S, Rzeszotarski J M. Designing interactive transfer learning tools for ML non-experts[C] //Proceedings of the CHI Conference on Human Factors in Computing Systems. 2021: 1-15.

[102] Cheng F, Keller M S, Qu H, et al. Polyphony: An Interactive

Transfer Learning Framework for Single-Cell Data Analysis[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(1): 591-601.

[103] Chen C, Wang Z, Wu J, et al. Interactive Graph Construction for Graph-based Semi-supervised Learning[J]. IEEE Transactions on Visualization and Computer Graphics, 2021, 27(9): 3701-3716.

[104] Chen C, Wu J, Wang X, et al. Towards Better Caption Supervision for Object Detection[J]. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(4): 1941-1954.

[105] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: Early experiments with gpt-4[J]. arXiv preprint arXiv:2303.12712, 2023.

[106] OpenAI. GPT-4 Technical Report [J]. arXiv preprint arXiv: 2303.08774, 2023.

[107] Liu H, Tam D, Muqeeth M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning[C] //Proceedings of the Advances in Neural Information Processing Systems, 2022: 1950-1965.

[108] Min S, Lyu X, Holtzman A, et al. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?[C].//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022: 11048–11064.

[109] Liu J, Shen D, Zhang Y, et al. What Makes Good In-Context Examples for GPT-3?[C]. //Proceedings of Deep Learning Inside Out: The Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, 2021: 100–114.

[110] Kim J, Kim H J, Cho H, et al. Ground-truth labels matter: A deeper look into input-label demonstrations[C] //Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022: 2422–2437.

[111] Lu Y, Bartolo M, Moore A, et al. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity[C] //Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2021: 8086–8098.

[112] Yang W, Wang X, Lu J, et al. Interactive steering of hierarchical clustering[J]. IEEE Transactions on Visualization and Computer Graphics, 2020, 27(10): 3953-3967.

作者姓名（按论文署名顺序填写）	单位	工作邮箱（域名为工作单位的邮箱）	特殊情况说明（包括单位邮箱涉密不可公开、非员工无单位邮箱等情况请说明）
杨维铠	清华大学软件学院	yangwk21@mails.tsinghua.edu.cn	联系电话:13051670559
陈长建	清华大学软件学院	ccj17@mails.tsinghua.edu.cn	
朱江宁	清华大学软件学院	zhu-jn19@mails.tsinghua.edu.cn	
李磊	中国航天科工集团第三研究院	univer1@sina.com	所在单位不提供单位邮箱
刘鹏	中国航天科工集团第三研究院	qiyanjie543@sina.com	所在单位不提供单位邮箱
刘世霞	清华大学软件学院	shixia@tsinghua.edu.cn	