Enhancing Training Data Quality With Visual Analytics

Shixia Liu, Tsinghua University

This installment of Computer's *series highlighting the work published in IEEE Computer Society journals comes from* IEEE Transactions on Visualization and Computer Graphics.

In this "Spotlight on Transactions" column, we take a closer look at selected research articles recently published in *IEEE Transactions on Visualization and Computer Graphics*. These articles explore visual analytics methods to identify and diagnose data-quality issues, such as out-of-distribution (OoD) instances, noisy annotations, and imprecise annotations. Given that data consist of individual instances and their associated annotations, visual analytics efforts for enhancing training data quality can be categorized into three main classes: instance diagnosis, annotation diagnosis, and hybrid diagnosis.

Training data quality is of paramount importance to the success of machine learning. No matter how complicated the learning model is and how much data are used to train it, the performance of the model is ultimately limited by the quality of the training data. However, obtaining high-quality training data is a time-consuming and expensive process. This has sparked a recent shift in artificial intelligence (AI) system development, moving from model-centric AI to data-centric AI. As a result, the field of visual analytics has witnessed growing research dedicated to enhancing training data quality. This research trend involves a tight integration of interactive visualization techniques with machine learning techniques, offering new and promising avenues to address the challenges posed by data-centric AI systems.

Instance diagnosis involves identifying and addressing quality issues at the instance level, which includes OoD instances, blurry images, and data heterogeneity. OoD instances refer to data instances that significantly differ from the distribution of the training data used to train a machine learning model. Figure 1 illustrates two such examples. Handling OoD instances is crucial in many real-world applications to ensure the model's robustness and generalization to unseen data. To better identify and understand such instances, OoDAnalyzer[1] integrates an ensemble-based OoD detection method and a grid visualization. By enlarging

Digital Object Identifier 10.1109/MC.2023.3304568
Date of current version: 18 October 2023

the feature set and the algorithm set, the developed detection method can achieve better detection performance. To understand why these instances are not covered by the training data, a similarity-preserving grid visualization is developed for exploring these OoD instances in context. To support real-time interaction, a kNN-based approximation is proposed to speed up the layout algorithm. Once OoD instances and underlying reasons have been identified, adding labeled instances accordingly to training data can usually improve the model performance.

Since the publication of OoDAnalyzer,[1] researchers have been actively exploring and advancing visual analytics techniques to enhance data quality at the instance level. For example, Wang et al.[2] developed a visual analytics method, HetVis, to address the data heterogeneity challenge in horizontal federated learning (HFL). Data heterogeneity among clients is a critical concern when training high-quality HFL models. HetVis enables participating clients to explore and understand data heterogeneity through the comparison of prediction behaviors between the global federated model and local stand-alone models. The key feature of this method lies in its capacity to employ context-aware clustering to summarize inconsistent records and provide visualizations to identify heterogeneity issues in HFL. This capability serves as a source of inspiration for improving the quality of local datasets.

Annotation diagnosis centers on evaluating and resolving issues related to annotations provided for instances. This includes identifying and addressing problems, such as inaccurate annotations, where the provided annotations for instances are incorrect; incomplete annotations, where essential annotations are missing, resulting in none or only a few instances being annotated; and inexact annotations, where the provided annotations are coarse-grained

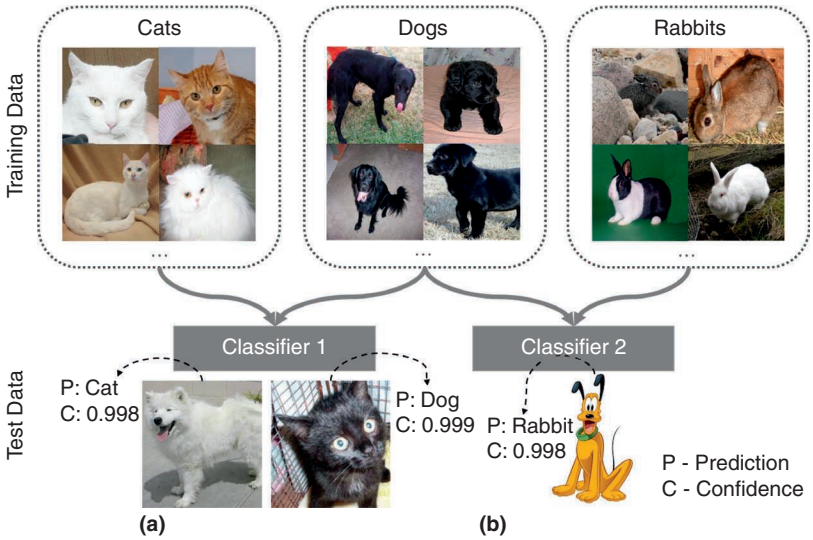

FIGURE 1. OoD samples in image classification. (a) A white dog and a black cat are incorrectly predicted with high confidence by a classifier trained on a dataset only consisting of dark-colored dogs and light-colored cats. (b) A more complex case where a cartoon dog with long ears is incorrectly predicted as a rabbit due to the absence of cartoon dog in training data.

and not as precise as required for the specific task. Addressing these annotation-related issues is crucial to ensure the reliability and generalizability of machine learning models and their subsequent applications in real-world scenarios. To address the issue of incomplete annotations, Park et al.[3] introduced CMed, a visual analytics framework tailored for exploring crowdsourced medical image data annotations. By integrating interactive linked visualizations, CMed allows detailed examination of crowd annotation results for specific videos and workers. This capability empowers users to observe patterns and gain valuable insights into crowdsourced data annotations, thereby facilitating the design of more effective crowdsourcing applications. CMed proves to be a useful tool in improving the overall quality and efficiency of the crowdsourcing process. Subsequently, FSLDiagnotor was introduced as a solution to address the challenges of inaccurate and incomplete annotations.[4] This method formulates the representative instance selection as a sparse subset selection problem. Based on this formulation, it effectively filters out low-quality annotated instances and suggests the addition of new instances for annotation to ensure a comprehensive representation of the data collection.

Hybrid diagnosis combines both instance-level and annotation-level assessments to identify and address quality issues in the data used for training machine learning models. Such a comprehensive diagnosis method is essential in many machine learning tasks. For example, evaluating models under different subsets of data, known as *data slices*, is essential to ensure fairness and consistent performance in diverse situations. In response to this requirement, Zhang et al.[5] developed SliceTeller, a tool designed to identify and address issues with critical data slices. It enables users to debug and enhance machine learning models by detecting problematic data slices. Once the underlying

issues are identified, users can either correct noisy annotations or adjust the weights of the associated instances. To speed up model iterations, this tool also employs a boosting model to estimate performance divergences after a slice-based model fine-tuning. Additionally, SliceTeller facilitates the comparison of multiple model versions, streamlining the selection of the most suitable one for a given application. Using this hybrid diagnosis, SliceTeller exemplifies the drive to improve the robustness and fairness of machine learning models.

The tight integration of visual analytics techniques with machine learning techniques has been particularly effective in enhancing the quality of training data. Given the promising results and trends observed, we expect even greater achievements from this synergy in the future. This is especially true in today's era, characterized by the prevalence of large foundation models, where the quality of data for fine-tuning these models is crucial for adapting them to specific tasks. As a premier source of knowledge, *IEEE Transactions on Visualization and Computer Graphics* remains committed to sharing pivotal research findings on this evolving topic.

REFERENCES

1. C. Chen et al., "OoDAnalyzer: Interactive analysis of out-of-distribution samples," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 7, pp. 3335–3349, Jul. 2021, doi: 10.1109/TVCG.2020.2973258.
2. X. Wang, W. Chen, J. Xia, Z. Wen, R. Zhu, and T. Schreck, "HetVis: A visual analysis approach for identifying data heterogeneity in horizontal federated learning," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 1, pp. 310–319, Jan. 2023, doi: 10.1109/TVCG.2022.3209347.
3. J. H. Park, S. Nadeem, S. Boorboor, J. Marino, and A. Kaufman, "CMed: Crowd analytics for medical imaging data," *IEEE Trans. Vis. Comput. Graphics*, vol. 27, no. 6, pp. 2869–2880, Jun. 2021, doi: 10.1109/TVCG.2019.2953026.
4. W. Yang et al., "Diagnosing ensemble few-shot classifiers," *IEEE Trans. Vis. Comput. Graphics*, vol. 28, no. 9, pp. 3292–3306, Sep. 2022, doi: 10.1109/TVCG.2022.3182488.
5. X. Zhang, J. P. Ono, H. Song, L. Gou, K.-L. Ma, and L. Ren, "SliceTeller: A data slice-driven approach for machine learning model validation," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 1, pp. 842–852, Jan. 2023, doi: 10.1109/TVCG.2022.3209465.

SHIXIA LIU is a professor in the School of Software at Tsinghua University, Beijing 100190, China. She is currently the associate editor in chief of *IEEE Transactions on Visualization and Computer Graphics*. Contact her at shixia@tsinghua.edu.cn.

Digital Object Identifier 10.1109/MC.2023.3317985