Interactive Visualization of Ensemble Decision Trees based on the Relations among Weak Learners

Miyu Kashiyama
Ochanomizu University
Tokyo, Japan
kashiyama.miyu@is.ocha.ac.jp

Masakazu Hirokawa
NEC Corporation
Tokyo, Japan
masakazu-hirokawa@nec.com

Ryuta Matsuno
NEC Corporation
Tokyo, Japan
ryuta-matsuno@nec.com

Keita Sakuma
NEC Corporation
Tokyo, Japan
keita.skm@nec.com

Takayuki Itoh
Ochanomizu University
Tokyo, Japan
itot@is.ocha.ac.jp

***Abstract*—Ensemble learning that combines multiple weak learners for enhanced performance, is widely used but suffers from low interpretability/explainability. This leads challenges not only in operational aspects like model maintenance and quality assurance but also in addressing societal needs such as fairness and privacy. To tackle this, we propose a new visualization method focusing on the relationship among weak learners in ensemble models to improve understanding of the model structure and its learning processes. In this paper, we defined the relation between weak learners based on a "common sample" in gradient-boosting decision trees, and a visualization method as a three-dimensional graph structure was proposed. Ensemble models trained with synthetic data sets that include typical distribution shifts and real-world open data sets were visualized. As a result, we demonstrated that this approach enables a more accessible understanding of the behavior and structure of ensemble models comprising multiple weak learners, facilitating the identification of overfitting and underfitting through visualization of changes during the training and validation processes.**

***Index Terms*—Visualization, Machine learning, Ensemble model**

I. Introduction

Predictive models using machine learning techniques are widely employed across various industries. Among machine learning methods, ensemble learning aggregates multiple weak learners to achieve robust expressive and predictive performance, rendering it widely favored for its practical implementation [1]. However, the complexity caused by the large ensemble size brings challenges in comprehending the overall model structure and in providing interpretability and explanatory insights to users. Analyzing the model structure is crucial not only for operational considerations such as model quality assurance and maintenance but also for increasing interpretability/explainability to address social imperatives such as fairness and privacy protection. Thus, various information visualization techniques have been proposed to facilitate the analysis and assessment of machine learning model structures. Meng et al. [2] proposed a visualization system that compares multiple models based on training data features and model performance. Similarly, Zhang et al. [3] and Nagasaka et al. [4] introduced visualization systems for deep learning models within immersive environments, demonstrating the effectiveness of visualization within a three-dimensional space.

This study proposes a novel visualization approach that allows users to graphically overview the model structure of ensemble decision trees based on the relationships among weak learners in the model using an interactive interface. We also provide interactive functions aimed at enhancing the interpretability and adjustability of developed models. This method facilitates comprehension of the behavior and structure of ensemble models comprising multiple weak learners. Moreover, by visualizing changes throughout training and validation, it makes users easier to identify model overfitting and underfitting, significant sources of error in machine learning applications. It helps users gain insights regarding the challenges in machine learning operations (MLOps [5]) such as evaluating the credibility of the learning outcome, detecting the distribution changes in the dataset, and interpreting the model output including identifying the cause of the prediction errors in the operation phase [6]. We adopted a gradient-boosting decision tree (GBDT) [7] as an exampled ensemble decision tree. This paper demonstrates the utility of the proposed visualization method through experiments with synthetic datasets and real-world open datasets,

II. Related Works

A. Visualization of Decision Trees

Numerous studies have been published on visualization techniques aimed at improving the interpretability and prediction accuracy of machine learning models. Wang et al. introduced an interactive visualization tool [8] designed to assist users in exploring models that align with their domain knowledge and values among a collection of sparse decision trees exhibiting similar performance but differing in data explanatory properties (known as "Rashomon" effect). Kovalerchuk et al. [9] proposed two novel decision tree visualization methods tailored for machine learning. This study demonstrated that observing and analyzing relationships

2375-0138/24/$31.00 ©2024 IEEE
DOI 10.1109/IV64223.2024.00028

among attributes, decision tree structure, and data flow within the decision tree can effectively mitigate over generalization and overfitting of models.

B. Visualization of Ensemble Learning

While traditional decision tree visualization approaches focused on representing the structure and data flow of an individual tree, several studies attempted to visualize ensemble model as a whole for improving its performance based on facilitating the understanding of entire model behavior. StackGenVis [1] is a system designed to assist users in dynamically adapting performance metrics, managing data instances, selecting essential features, and identifying effective algorithms within a given dataset to measure predictive performance. The system enables the discovery of not only the performance of the final model but also the identification of over-promising or under-performing models utilized throughout the process. Nsch et al. [10] proposed a visualization approach inspired by botanical principles that enhances intuitive comprehension within a two-dimensional framework. This method enables tailored representation of random forest characteristics through visual attributes, facilitating interactive analysis of forest structure. However, it is still challenging in revealing continuous relationships among weak learners, such as those in GBDT, using these approaches. Moreover, a small number of studies have focused on elucidating the collective properties of entire ensemble models stemming from interactions among weak learners.

Regarding the study of visualization of the GBDT model, a system focused on advertising datasets was proposed. Gao et al. [11] proposed a visual analysis system to help advertising analysts understand the working mechanism of GBDT-based CTR prediction models from three levels: instance, feature, and model, and to facilitate the model tuning process. Based on the background of online advertising, combined with the three key participants in online advertising campaigns and the CTR prediction model building and tuning process, analysts can intuitively explore instances of advertising data and the corresponding prediction results.

This study proposes a novel visualization method tailored for ensemble decision trees on any dataset, addressing the aforementioned limitations. The focus lies on the concurrent examination of not only predictive data features and model performance outcomes but also delineating interrelationships among weak learners (hereafter, "trees") and discerning how variations in dataset composition influence the learning history of the models. The proposed approach aims to foster a shared understanding and collaborative exploration between machine learning practitioners and domain experts responsible for model decision-making processes.

III. Proposed Method

A. Overview of the Ensemble Model Visualization

The following is an overview of the proposed ensemble model visualization method. As shown in Fig. 1, the method visualizes a model structure of GBDT as a three-dimensional

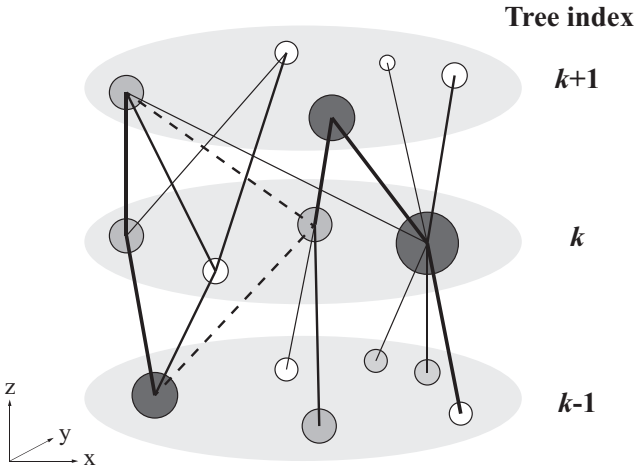

Fig. 1. Conceptual diagram of the proposed visualization for ensemble decision tree.

plot of a layered graph structure along the z-axis, where each layer that lies in the xy-plane represents a tree in the GBDT and each node represents a leaf node in the tree. The size and color of a node represent the number of samples of a data set allocated to the leaf node and the average prediction errors, respectively. While the distance between two nodes in the same layer represents the similarity of features of the samples allocated in the corresponding leaves in the trees, the distance between nodes in two adjacent layers represents the similarity of the features as well as the commonality of the samples in the leaves. The solid links connecting nodes in the figure visualize this commonality among training data, while dashed links highlight the commonalities that appeared at inference (i.e., among test data) that did not appear at the training time. The following subsection explains the details.

1) Relations Among Weak Learners: Our visualization aims at providing insights regarding the relations among the weak learners, i.e., decision trees, consisting of ensemble trees. Specifically, we utilize the commonality of the samples allocated in the leaf nodes. The commonality of a leaf i of the kth tree and a leaf j of the $(k+1)$th tree is defined as the ratio of the samples commonly allocated to the leaves i and j among the sample allocated to the leaf i. More formally, the ratio of the common samples $c_{i,j}^{(k)}$ is defined as:

$$c_{i,j}^{(k)} = \frac{\left| S_i^{(k)} \cup S_j^{(k+1)} \right|}{\left| S_i^{(k)} \right|}, \tag{1}$$

where $S_i^{(k)}$ is the set of samples allocated to the leaf i in the kth tree among a dataset. The thickness of the solid links in Fig. 1 is based on the ratio of the common samples.

2) Optimization of Node Coordinates: The coordinates of each node on the xy-plane are computed as follows:

1) For each the kth decision tree, compute initial embedding matrix $V^{(k)} = [v_1^{(k)}, ..., v_{n_k}^{(k)}]^\mathsf{T}$, where n_k is the

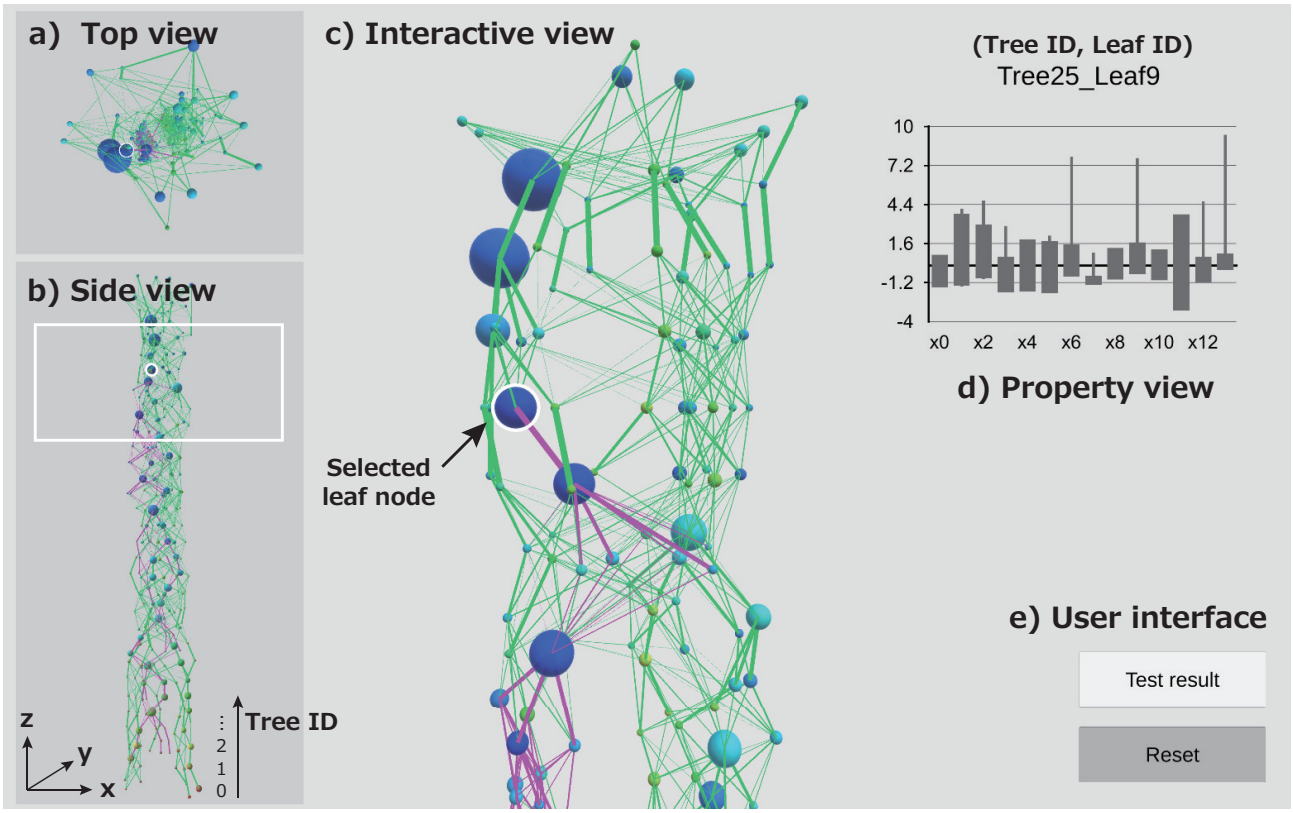

Fig. 2. Overview of the proposed visualization method. This consists of: a) top view, b) side view, c) interactive view, d) property view, and e) user interface. The interactive view shows the specified area indicated as the white rectangle in the side view, and the detailed information (e.g., feature distributions) of the selected leaf node is displayed in the property view.

number of samples in the tree, based on the distance matrix $D^{(k)}$, where $D^{(k)}_{ij} := \|u^{(k)}_i - u^{(k)}_j\|_2$ and $u^{(k)}_i$ is the average feature of the samples in $S^{(k)}_{ij}$.

2) To place similar nodes close in our visualization, minimize the following loss function by gradient descent w.r.t. V for 500 epochs:

$$\sum_k \left\| \widehat{D}^{(k)} - D^{(k)} \right\|_2 + \sum_{i,j} c^{(k)}_{i,j} \left\| v^{(k)}_i - v^{(k+1)}_j \right\|_2 \quad (2)$$

where $\widehat{D}^{(k)}_{ij} := \|v^{(k)}_i - v^{(k)}_j\|_2$. The first term considers the nodes in each layer while the second term is designed for the nodes in two adjacent layers.

3) To avoid local optimal solutions due to "twisting" of the links, for every 100 epochs in the previous step, we rotate each $V^{(k)}$ around the origin $(0, 0)$ so that the second term in 2 is minimized.[1]

B. Interactive Interface for Proposed Visualization

Fig. 2 shows the prototype interface of the proposed visualization method implemented using Unity[2]. This consists of: a) top view, b) side view, c) interactive view, d) property view, and e) user interface. In the interactive view, a user can view the visualized model while changing the viewing angle and scale interactively. The specified area being viewed in the interactive view was indicated as the white rectangle in the side view. When a user selectes a leaf node in the interactive view, the detailed information (e.g., feature distributions of samples contained in the node) is displayed in the property view, and the likes connected to the selected node were highlighted in purple.

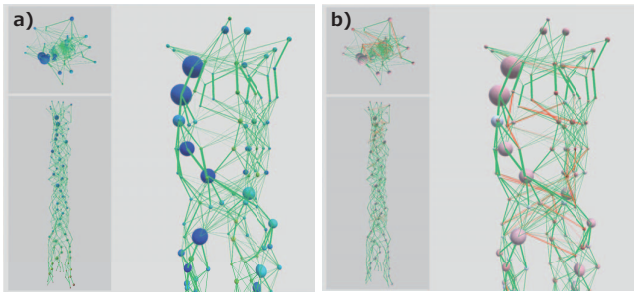

Fig. 3. Differential view between two datasets. a) The result of training data, and b) that of test data. In b) node color denotes the difference of residual and the red links indicate novel links compared to the training data.

[1]Mirrored rotations are also considered.
[2]https://unity.com

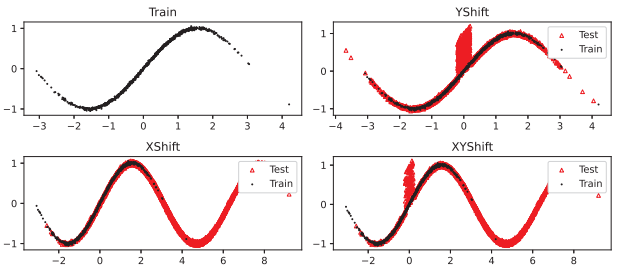

Fig. 4. Examples of distribution changes with artificially created datasets. Top left) Training dataset, Top right, Bottom left, and Bottom right) Test dataset with Y-shift, X-shift, and XY-shift, respectively.

Moreover, upon activating the Test Result button, the difference between the training and test data sets is visually depicted as *differential view* shown in Fig. 3. Fig. 3a) shows the visualization with the training data, and Fig. 3b) shows the model prediction result when the test data is fed to the trained model. Changes in error magnitude between these datasets are indicated by node color variations; nodes turning red denote increased errors relative to the training data, while blue nodes denote error reductions. Links highlighted in red color depict common samples that did not appear in the training data as explained in Section III-A. This feature offers a holistic overview of the model training outcomes, facilitating the identification of anomalous leaf nodes and enhancing comprehension of dataset disparities. Additionally, this feature provides detailed insights into individual trees and pertinent data samples to aid in model refinement and the analysis of predictive error causes.

IV. Experiments

A. Visualizing Distribution Changes with Synthetic Dataset

This section verifies how the proposed method visualizes diverse distributional shifts between the dataset employed for training the ensemble model and the dataset presumed to manifest post-operation. Specifically, we generated the following three datasets synthetically to illustrate input-output relationships within the training one-dimensional regression dataset depicted in Fig. 4 (upper left).

- Partial Distributional changes in the objective variable (*Y-shift* in Fig. 4 (upper left))
- Distributional changes in explanatory variables(*X-shift* in Fig. 4 (lower left))
- Distributional changes in both the objective and explanatory variables (*XY-shift* in Fig. 4 (lower right))

Fig. 5 summarizes the visualization outcomes of a GBDT model (specifically LightGBM [12]) constructed with the synthetic datasets. Fig. 5a) visualizes the result of the training data, and b), c), and d) show the differential view of *Y-shift*, *X-shift*, and *XY-shift* result from the training data, respectively. Observing Fig. 5b) reveals that alterations in the distribution of the target variable correspond to increased output errors

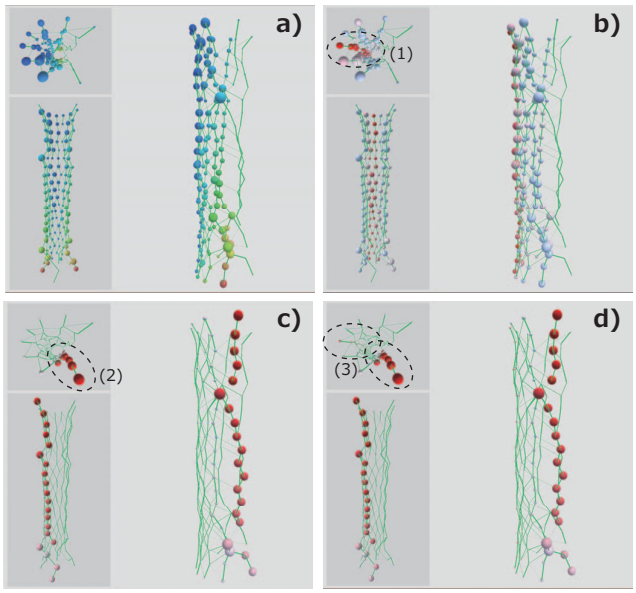

Fig. 5. Visualization of the synthetic datasets with distribution shifts. a) Training dataset, b) Y-shift, c) X-shift, and d) XY-shift. b), c), and d) are shown as differential views with a).

TABLE I
Real-world open datasets used

Dataset	Samples	Features
Breast Cancer	683	10
Australian Credit Approval	690	14
Cpusmall	8192	12

at particular nodes indicated with the dashed ellipse (1). Likewise, shifts in the explanatory variable distribution manifest as notable changes in the data sample count as size variations of the nodes, particularly in regions with smaller sample sizes during training. Since those samples were not included in the training data, the prediction error of these nodes (dashed ellipse (2) in Fig. 5c)) significantly increased indicated by the deep red color of the node. Moreover, when both the target and explanatory variable distributions change, these alterations are concurrently visualized through node color and size difference (dashed ellipses (3) in Fig. 5d)).

B. Case Studies with Real-World Datasets

Subsequently, we visualized a GBDT model (LightGBM) trained on real-world data summarized in Table I for a publicly accessible regression task [13] using the proposed method. Herein, we present the estimation outcomes derived from the GBDT model comprising 30 trees, applied to mixed data.

1) Breast Cancer Dataset: Fig. 6 depicts the visualization outcomes. The figure on the right illustrates the differential view of the test dataset compared to the training dataset on its left. As shown in the figure, many novel common samples appeared in the test dataset (indicated as red dashed links) and the prediction errors of nodes containing those samples

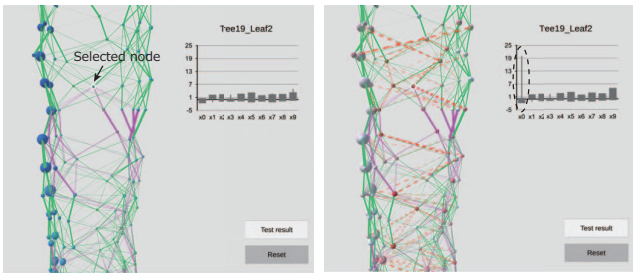

Fig. 6. Differential view in Breast Cancer dataset between training data (left) and test data (right).

increased. The feature distributions of the same node are shown in the property view in both figures. From this view, we could observe that samples the value of feature x_0 significantly varied from that of training data contained in the test data, and hence it influenced the prediction error. Consequently, we supposed that the absence of samples from regions close to the test data in the training dataset led to overfitting and an increase in node errors.

2) Australian Credit Approval Dataset: Fig. 7a), b), and c) shows the visualization result (top/side views) of the Australian dataset with 10, 20, and 30 trees, respectively. The model structure was initially separated into two pathways as shown in Fig.7a)b), and subsequently converged around the 23rd tree, followed by branching again as highlighted in Fig. 7d). This visualization result suggests a substantial dichotomy in the feature space of samples within the training data. Fig. 8 presents a conventional visualization of decision trees for the 0th (top row), 20th (2nd row), 23rd (3rd row), and 25th (bottom row) trees. Notably, for the 23rd tree, we observed a significant alteration in the splitting criteria from previous trees, such as the 0th and 20th trees. Furthermore, in Fig. 7 (d), the trajectory of sample paths within nodes after the merging and re-branching after the 23rd tree reveals a return to a previously encountered region. This was also observed as the similar root splitting criteria of the 25th tree to that of the 0th/20th tree, as shown in the bottom row of Fig. 8. This behavior suggests that a strategy of data splitting before training is likely to yield better outcomes.

3) Cpusmall Dataset: Fig. 9 depicts the visualization results of the Cpusmall dataset in the top view (left) and side view (right). We found a more pronounced separation in the structure in this figure compared to the visualization of the Australian dataset. In this case, the majority of samples concentrated in the region indicated in Fig. 9(1), whereas certain samples followed a clearly separated route as the highlighted links in Fig. 9(2). Fig. 10 shows the target value distribution of samples in the dataset. The distribution colored in blue represents the target values of samples in the region (1), while the orange distribution represents that of samples in the region (2). This implies that the samples in the region (2) hold not only variated feature values from that in the region (1) but also isolated target values. The observation of such

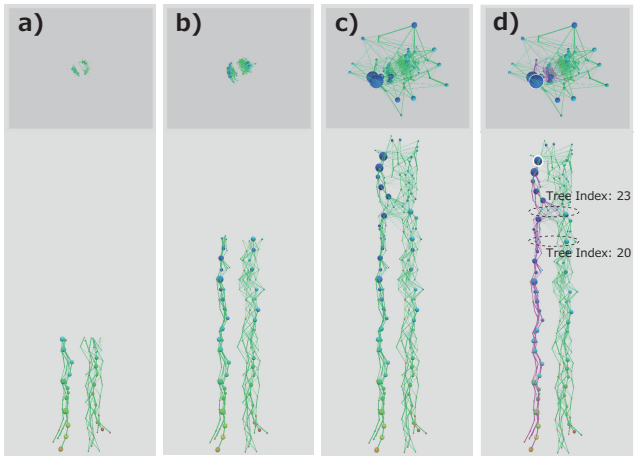

Fig. 7. Visualization of Australian dataset with the different number of trees. The number of trees was set as: a) 10, b) 20, and c) 30, respectively. d) Highlighting the route to the selected node.

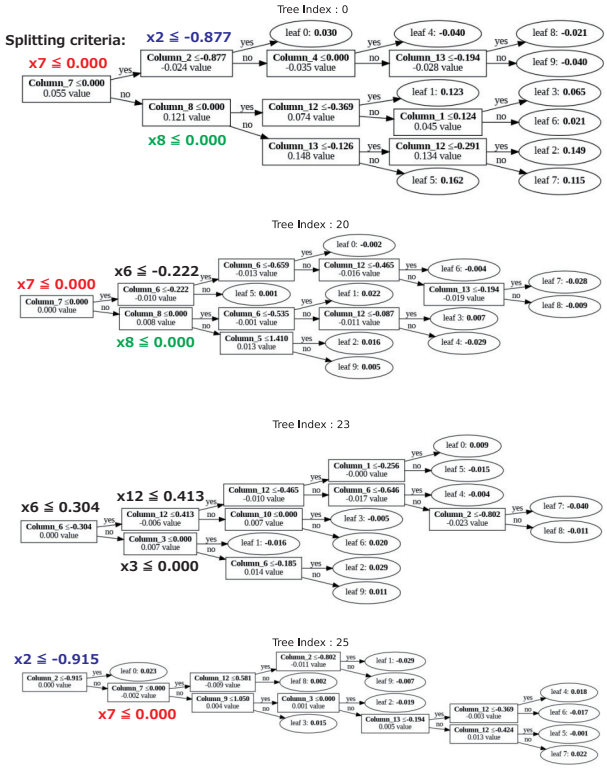

Fig. 8. Tree diagram of 0 th (top row), 20 th (2nd row), 23 rd (3rd row), and 25 th (bottom row) tree. The same/similar root-splitting rules are highlighted with the common text color.

a unique route suggests the possibility of identifying outlier samples within the dataset and/or evaluating if the trained model successfully splits the problem into subproblems.

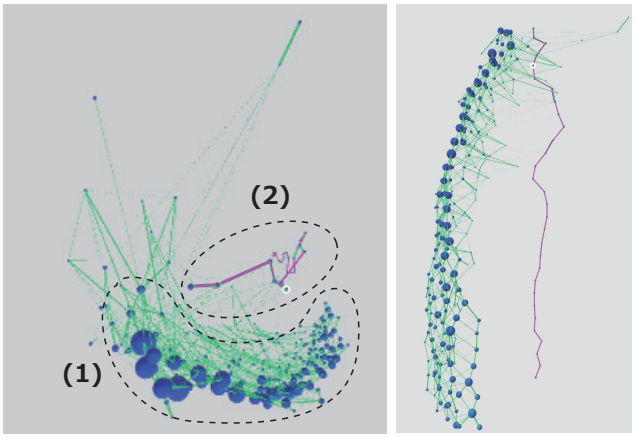

Fig. 9. Visualization of Cpusmall dataset, left) top view, and right) side view.

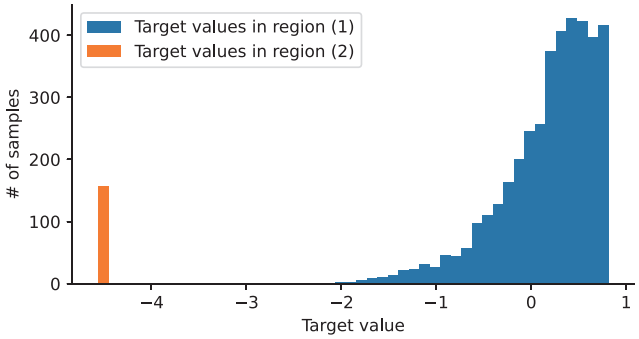

Fig. 10. Target value distribution of samples in two regions is shown in Fig. 9.

Conclusion

In this paper, we introduced a novel visualization method that emphasizes the relationships among weak learners, presenting them as a graph structure in three-dimensional space to facilitate comprehension of the behavior and structure of ensemble models comprising multiple weak learners. The proposed method enables the presentation of typical distributional changes in a dataset—critical issues in the practical operation of machine learning models in a visually accessible format. This visualization aids in identifying instances of overfitting and underfitting, challenges often not readily discernible through conventional methods. Additionally, the method assists in determining the potential for problem decomposition into subproblems when root bifurcations occur within the model structure. Subproblem decomposition in Decision Trees is vital for enhancing accuracy and interpretability, as emphasized in prior studies [14] [15]. While the proposed method does not directly partition into subproblems, it provides data scientists with intuitive insights into factors contributing to increased prediction error and potential solutions, offering valuable guidance in problem-solving contexts. As future work, we aim to validate the efficacy of our visualization tool as an analytical instrument for both the development and operation of complex, large-scale machine learning systems. Furthermore, we intend to apply this method to advance technologies within the domain of MLOps.

References

[1] A. Chatzimparmpas, R. M. Martins, K. Kucher, and A. Kerren, "Stackgenvis: Alignment of data, algorithms, and models for stacking ensemble learning using performance metrics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1547–1557, 2021.

[2] L. Meng, S. van den Elzen, and A. Vilanova, "Modelwise: Interactive model comparison for model diagnosis, improvement and selection," *Computer Graphics Forum*, vol. 41, no. 3, pp. 97–108, 2022.

[3] X. Zhang, Z. Yin, Y. Feng, Q. Shi, J. Liu, and Z. Chen, "Neuralvis: Visualizing and interpreting deep learning models," in *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 1106–1109, 2019.

[4] H. Nagasaka and M. Izuhara, "Interactive visualization of deep learning models in an immersive environment," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21, (New York, NY, USA), Association for Computing Machinery, 2021.

[5] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine learning operations (mlops): Overview, definition, and architecture," *IEEE Access*, vol. 11, pp. 31866–31879, 2022.

[6] K. Sakuma, R. Matsuno, and Y. Kameda, "A method of identifying causes of prediction errors to accelerate mlops," in *2023 IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest)*, pp. 9–16, 2023.

[7] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001.

[8] Z. J. Wang, C. Zhong, R. Xin, T. Takagi, Z. Chen, D. H. Chau, C. Rudin, and M. Seltzer, "Timbertrek: Exploring and curating sparse decision trees with interactive visualization," in *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 60–64, 2022.

[9] B. K. A. Dunn, A. Worland, and S. Wagle, "Interactive decision tree creation and enhancement with complete visualization for explainable modeling," 2023.

[10] R. H. Nsch, P. Wiesner, S. Wendler, and O. Hellwich, "Colorful trees: Visualizing random forests for analysis and interpretation," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 294–302, 2019.

[11] W. Gao, S. Liu, Y. Zhou, F. Wang, F. Zhou, and M. Zhu, "Gbdt4ctrvis: visual analytics of gradient boosting decision tree for advertisement click-through rate prediction," *Journal of Visualization*, pp. 1–21, 2024.

[12] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 3149–3157, Curran Associates Inc., 2017.

[13] R.-E. Fan, "Libsvm data: Classification, regression, and multi-label." https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets.

[14] J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer, "Generalized and scalable optimal sparse decision trees," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 6150–6160, PMLR, 13–18 Jul 2020.

[15] H. McTavish, C. Zhong, R. Achermann, I. Karimalis, J. Chen, C. Rudin, and M. Seltzer, "Fast sparse decision tree optimization via reference ensembles," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 9604–9613, Jun. 2022.