Model-Agnostic Visual Explanation of Machine Learning Models Based on Heat Map

S. Sawada[1] and M. Toyoda[2]

[1]The University of Tokyo, Japan
[2]Institute of Industrial Science, The University of Tokyo

Abstract

It is essential to assess the trustworthiness of the machine learning models when deploying them to real-world applications, such as healthcare and risk management, in which domain experts need to make critical decisions. We propose a visual analysis method for supporting domain experts to understand and improve a given machine learning model based on a model-agnostic interpretable explanation technique. Our visualization method provides a heat map matrix as an overview of the model explanation and helps efficient feature engineering and data cleaning. We demonstrate our visualization method on a text classification task.

CCS Concepts

• ***Human-centered computing*** → ***Heat maps; Visual analytics;*** • ***Computing methodologies*** → *Machine learning;*

1. Introduction

In this paper, we propose a visual analysis method for understanding machine learning models and improving them by feature engineering and data cleaning. While sophisticated machine learning models achieve high accuracy in various tasks, the need for interpretable explanation of the models is growing in application domains such as healthcare, and risk management.

To help domain experts understand a given trained model, we adopt an explanation technique that locally explains the prediction of each instance, such as LIME [RSG16] and SHAP [LL17]. Our visualization method provides an overview of a large set of local explanations by a heat map matrix that is grouped by similar instances and features. It enables domain experts to identify sets of features and instances to be modified for improving the model or cleaning the dataset.

2. Related Work

There have been several studies on visual analytic methods for understanding and refining machine learning models through their explanation. There are two types of such methods: white-box visual explanation and black-box visual explanation.

White-box visual explanation approaches [WSW*18] [SGPR18] [WGSY19] try to directly visualize the behavior of the models especially for machine learning experts. In contrast, black-box visual explanation approaches apply model-agnostic explanation techniques that separately develop interpretable models approximating the original models.

The black-box approach can be applied to any models, and the explanation can be a decision tree [CS96], a linear prediction model [RSG16], and a rule-based model [RSG18]. Krause et al. [KDS*17] utilized an explanation with a set of features that are prominent for predicting an instance and developed a visual analytic workflow. RuleMatrix [MQB19] visualizes rule-based explanations of an entire model. Some studies [ZWM*19] [KCK*19] provide visualization of statistical data of the model as explanations.

Our method adopts the black-box approach for supporting domain experts, and provides a novel heat map matrix based visualization with co-clustering of features and instances for efficient model improvement and data cleaning.

3. Explanation Visualization by Clustered Heat Map Matrix

Here, we introduce our visualization method based on clustered heat map matrix. Figure 1 shows the processing flow of our method. Given a trained machine learning model, it first generates a set

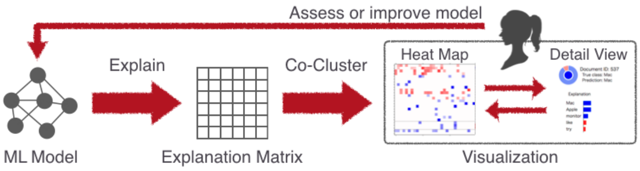
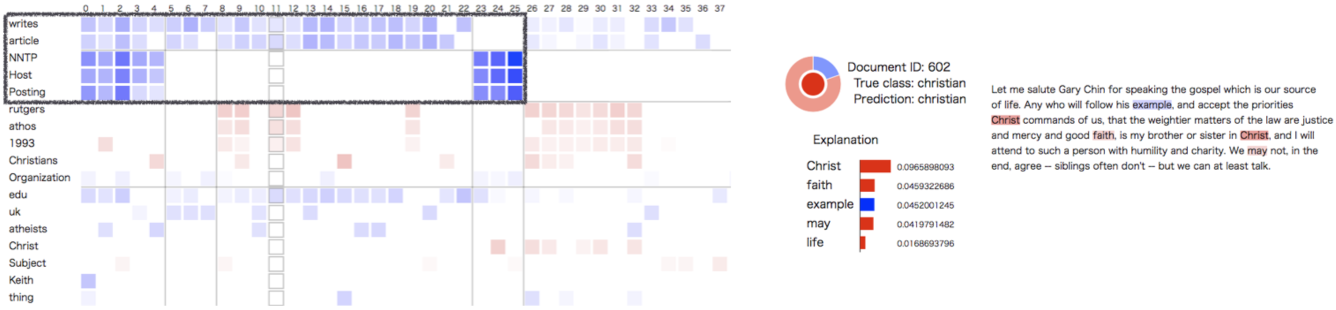


Figure 1: *The processing flow for creating our visual explanation.*

DOI: 10.2312/eurp.20191140



Let me salute Gary Chin for speaking the gospel which is our source of life. Any who will follow his example, and accept the priorities Christ commands of us, that the weightier matters of the law are justice and mercy and good faith, is my brother or sister in Christ, and I will attend to such a person with humility and charity. We may not, in the end, agree – siblings often don't – but we can at least talk.

Figure 2: *Our visualization consists of two views. The left-hand side is a heat map matrix as an overview of a large set of local explanations. The right-hand side is a detail view to display information and the local explanation of the designated instance.*

of interpretable explanations for individual predictions. Currently, we utilize the LIME [RSG16] algorithm for generating individual explanations consisting of a set of features and their importance weights. Given the prediction of an instance, LIME builds a linear prediction model that locally approximates the original model around the instance, and then its K or fewer features and their weights are regarded as the local explanation. For selecting instances to visualize, we utilize SP-LIME that is an extension of LIME for analyzing the entire model. The explanation generated by SP-LIME is a set of the explanations of B instances, which are selected so that they can cover important features as many as possible. Note that the importance of the feature is a sum of its weights, and the importance of a instance is a sum of the importance of features included in SP-LIME. Our method visualize this SP-LIME explanation as an $F \times B$ matrix (F is the total number of features in the SP-LIME explanation) with features as rows and instances as columns, and call it an explanation matrix.

Second, our method groups similar features and instances in the explanation matrix, so that the user can easily locate sets of features and instances that affect the accuracy of the original model. For clustering, we utilize a co-clustering method that performs simultaneous clustering on the row and column dimension of the matrix [MO04]. We sort features and instances inside each cluster according to their importance, and sort clusters according to the average importance of the elements of clusters. Before clustering, we remove features whose importance values are below a threshold because the explanation matrix is usually sparse and they could be noisy.

Then we visualize the clustered and sorted explanation matrix by a heat map (Figure 2 left). Each column indicates the local explanation of the instances. Cells with positive weights are colored blue, and those with negative weights are colored red. The clusters are divided by lines, so that the user can easily see sets of features and instances.

The right-hand side of Figure 2 is the detailed view which shows information and local explanation of designated instance. Colors used in this view correspond to the heat map. The pie chart shows the prediction result of the selected instance by the original model, and the color of the circle inside the pie chart represents the ground truth. The explanation of this instance by LIME, features and weights as a bar chart. The original instance is presented in the right above area. In Figure 2, since we supposed to document classification task, the document with important features (highlighted by colors) is displayed.

4. Case Study

In this section, we demonstrate an analysis of an machine learning model by our visualization method through a text classification task. We used part of the 20 newsgroups text dataset in scikit-learn(`http://scikit-learn.org/0.19/datasets/twenty_newsgroups.html`) including documents classified into "Atheism" and "Christianity". These documents are mail data including headers and footers. The number of training data and test data are 1079 and 717 respectively. In this study, Atheism corresponds to positive and Christianity corresponds to negative. We set a limit of the number of features (K) to 20, and the number of instances (B) to 50. We trained a random forest with 500 trees as the original model.

The original model achieved an accuracy score of 91.5%. Figure 2 (left) visualizes the explanation matrix of this model. We found that several clusters emphasized by the gray rectangle in Figure 2 affect a bunch of instances to be classified as Atheism. Most features in these clusters are mail headers which have nothing to do with religion, so we can assess this model is untrustworthy because it learned inappropriate features even if the accuracy score is high.

Next, we attempted to improve this untrustworthy model. Since we found out that the mail headers cause the problem, we remove headers from each document, and then train the model again using the cleaned data. After this manipulation, though the accuracy dropped to 76.1%, important features appeared at the top of the explanation changed to religious terms with negative weights, such as God, Jesus, church. To confirm the trustworthyness of this model, we inspected instances explained by religious features on the detailed view in Figure 2 (right). The input document with highlights describes Christian words are properly weighted. Consequently, we can see this model became more trustworthy by removing noisy features.

References

[CS96] Craven M., Shavlik J. W.: Extracting Tree-Structured Representations of Trained Networks. In *Advances in Neural Information Processing Systems 8*, Touretzky D. S., Mozer M. C., Hasselmo M. E., (Eds.). MIT Press, 1996, pp. 24–30. 1

[KCK*19] Kwon B. C., Choi M., Kim J. T., Choi E., Kim Y. B., Kwon S., Sun J., Choo J.: RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 299–309. doi:10.1109/TVCG.2018.2865027. 1

[KDS*17] Krause J., Dasgupta A., Swartz J., Aphinyanaphongs Y., Bertini E.: A Workflow for Visual Diagnostics of Binary Classifiers using Instance-Level Explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)* (Phoenix, AZ, Oct. 2017), IEEE, pp. 162–172. doi:10.1109/VAST.2017.8585720. 1

[LL17] Lundberg S. M., Lee S.-I.: A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017, pp. 4765–4774. 1

[MO04] Madeira S. C., Oliveira A. L.: Biclustering algorithms for biological data analysis: A survey. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* (2004), 24–45. 2

[MQB19] Ming Y., Qu H., Bertini E.: RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 342–352. doi:10.1109/TVCG.2018.2864812. 1

[RSG16] Ribeiro M. T., Singh S., Guestrin C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, ACM, pp. 1135–1144. doi:10.1145/2939672.2939778. 1, 2

[RSG18] Ribeiro M. T., Singh S., Guestrin C.: Anchors: High-Precision Model-Agnostic Explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence* (Apr. 2018). 1

[SGPR18] Strobelt H., Gehrmann S., Pfister H., Rush A. M.: LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 667–676. doi:10.1109/TVCG.2017.2744158. 1

[WGSY19] Wang J., Gou L., Shen H., Yang H.: DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 288–298. doi:10.1109/TVCG.2018.2864504. 1

[WSW*18] Wongsuphasawat K., Smilkov D., Wexler J., Wilson J., Mané D., Fritz D., Krishnan D., Viégas F. B., Wattenberg M.: Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 1–12. doi:10.1109/TVCG.2017.2744878. 1

[ZWM*19] Zhang J., Wang Y., Molino P., Li L., Ebert D. S.: Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics 25*, 1 (Jan. 2019), 364–373. doi:10.1109/TVCG.2018.2864499. 1