Visualizing Machine Learning Through Dynamics

[1] Gaddam Keerthi, [2] Tirunagiri Kaushik Prasad, [3] Sudersan Behera
[1][2] B. Tech, Department of CSE-AIML, Sphoorthy Engineering College, Hyderabad, India
[3] Associate Professor, Department of CSE-AIML, Sphoorthy Engineering College, Hyderabad, India

***Abstract*—To train algorithm-specific models, machine learning algorithms and conventional data mining techniques typically need a lot of data, with little to no user input during the model-building phase. Sometimes, such a "big data'' based machine learning technique is impractical for use in settings like clinical trials where gathering or processing data is exceedingly costly or challenging. Furthermore, in some subjects, like the biological sciences, expert knowledge can be quite helpful when developing models. We present a novel technique for interactive machine learning and visual data mining using visual analytics in this research. This method uses multidimensional data visualization approaches to make it easier for users to participate in the mining and machine learning processes. This enhances the effectiveness of model creation by enabling dynamic user feedback in many ways, including data selection, data labeling, and data correction. This method can have a major influence on applications where obtaining huge amounts of data is difficult, as it can drastically reduce the amount of data needed to train an appropriate model. Two application problems—the handwriting recognition problem and the human cognitive score prediction problem—are used to evaluate the suggested methodology. The results of both experiments demonstrate that interactive machine learning and data mining aided by visualization can get the same accuracy as an automated process using far smaller training data sets.**

***Index Terms*— visual analytics, machine learning, data quality, multi-dimensional data visualization, user interaction.**

INTRODUCTION

(Huang Li, 2019)In machine learning, visualization, especially multi-dimensional data visualization, has become more and more crucial. The area of visual analytics was founded as a result of this shift in visualization from data viewing to an integral component of the analysis process [1]. Through interactive exploration and visual transformations, well-crafted visualization methods can successfully "decode" the information found in the data via visual analytics. In recent years, numerous effective uses of visual analytics have been documented in a variety of fields, including bioinformatics, medicine,
engineering, and social science. However, throughout the last ten years, automated data mining and data analytics have advanced significantly. One of the latest challenges in big data research is the successful integration of machine learning and data mining with visualization.

Data is used by machine learning methods, such as neural networks and support vector machines, to create computational models, which are high-dimensional spatial representations of nonlinear surfaces. After training, the models can be applied to various analysis tasks like regressions, predictions, and classifications. Machine learning has been significantly strengthened as an efficient solution to a wide range of big data processing issues by recent advancements in deep learning. Machine learning algorithms are automatic techniques that mostly function as "black boxes," meaning that users have very little knowledge of how and why the algorithms succeed or fail.

Although the primary goal of the underlying machine learning models is to facilitate data-driven learning, consumers may find them difficult to comprehend or utilize. The goal of interactive machine learning is to give users a way to comprehend and engage with the learning process through visualization [2]. It has several significant potential advantages:

1)Understanding

Without a thorough grasp of the many components of machine learning algorithms and how and why they operate, it can be challenging to increase the algorithms' performance and efficiency. In deep learning, where there are numerous layers and interrelated components, this is even more true.

2)Knowledge input

The performance of machine learning algorithms can be greatly enhanced with human knowledge input, especially in fields requiring specialized knowledge like

science, engineering, and medicine. Additionally, human instinct based on visual perception can beat computer algorithms. To enable user input into the machine learning system, such as feature selection, dimension reduction, parameter setting, or addition/revision of rules and associations, it is crucial to create a user feedback platform that supports visualization.

3) Reduction of Data

A substantial amount of data is typically needed for machine learning to train the computational model. In situations such as clinical trials, where data gathering, labeling, or processing is highly costly or challenging, this approach may not be feasible. The user can iteratively choose the most important and valuable subset of data to be added to the training process through interactive visualization of the machine learning process, making the process of developing models more data-efficient.

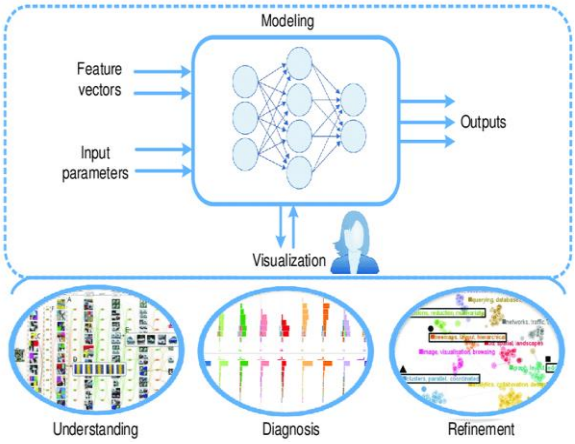

Fig. Overview of the visual analysis model

RELATED WORK

Although interactive machine learning has been previously proposed in the machine learning and AI communities [2, 3], applying visualization and visual analytics principles in interactive machine learning has only been an active research topic in recent years. Most of the existing studies focus on using visualization for a better understanding of machine learning algorithms. There have also been some recent works on using visual analytics for improving the performance of machine learning algorithms through better feature selection or parameter setting.

While there has been much literature on using interactive visualization to directly accomplish analysis tasks such as classification and regression [4, 5], we will focus mostly on approaches that deal with some machine learning models [6]. Previous works on using visualization to help understand the machine learning processes are usually designed for specific types of algorithms, for example, support vector machines, neural networks, and deep learning neural networks.

Neural Networks received the most attention due to the "black box" nature of the learning model and the complexity of its internal components. Multi-dimensional visualization techniques such as scatterplot matrices have been used to depict the relationships between different components of the neural networks [7, 8]. Typically, a learned component is represented as a higher dimensional point. The 2D projections of these points in either principal component analysis (PCA) spaces or a multi-dimensional scaling (MDS) space can better reveal the relationships of these components that are not easily understood, such as clusters and outliers. Several techniques have applied graph visualization techniques to visualize the topological structures of the neural networks [9, 10, 11]. Visual attributes of the graph can be used to represent various properties of the neural network models and processes.

Several recent studies tackle specific challenges in the visualization of deep neural networks due to the large number of components, connections, and layers. In [12]. Liu et al. developed a visual analytics system, CNNV, that helps machine learning experts understand deep convolutional neural networks by clustering the layers and neurons. Edge bundling is also used to reduce visual clutter. Techniques have also been developed to visualize the response of a deep neural network to a specific input in a real-time dynamic fashion [13, 14]. Observing the live activations that change in response to user input helps build valuable intuitions about how convents work.

PROPOSED SYSTEM

Our objective is to create a visualization-enabled platform for user interaction within a machine-learning context, enabling the user to track the development and functionality of the model's internal structures and offer suggestions that could enhance the algorithm's efficiency or reroute the model-building process. In this study, we mainly focus on "data reduction," however the visualization platform we design can be utilized to assist "understanding" and "knowledge input" activities as well. With our method, the user will be able to

pinpoint certain locations (in a certain visual space) where further information is required to enhance or rectify the model through the interactive system. In this manner, only the data required to learn a model is used. Our goal is to apply a little data solution to a big data problem. As the current, somewhat brute-force approach might not be necessary with smaller and higher quality data, in practice this approach can not only save costs for data acquisitions/collections in applications such as clinical trials, medical analyses, and environmental studies, but it can also improve the efficiency and robustness of machine learning algorithms.

To accomplish this, we must overcome the next two obstacles:

1. It is theoretically difficult to visualize a machine learning model's dynamics. Previous research frequently relies on certain machine learning techniques. However, we shall create a method and a broad approach in this study that works with the majority of machine learning algorithms. Support vector machines will be utilized as an example in our test applications to show how effective this method is.
2. It can be difficult to determine which parts of the visualization should be revised in the model and how best to quickly and effectively give the algorithm user feedback.
3. Machine learning features are frequently non-trivial data qualities that make it difficult to apply them in real-world applications to pre-screen possible targets for data collection.

We shall address a solution to these three issues in this study. We will evaluate our method on two real-world applications using datasets from the actual world.

IMPLEMENTATION

Machine learning models are becoming increasingly complex, powerful, and able to make accurate predictions. However, as these models become "black boxes," it's even harder to understand how they arrived at those predictions. This has led to a growing focus on machine learning interpretability and explainability.

For example, you applied for a loan at a bank but were rejected. You want to know the reason for the rejection, but the customer service agent responds that an algorithm dismissed the application, and they cannot determine the reason why. This is frustrating, right? You deserve an explanation for the decision that affects you. That's why companies try to make their machine-learning models more transparent and understandable.

One of the most promising tools for this process is SHAP values, which measure how much each feature (such as income, age, credit score, etc.) contributes to the model's prediction. SHAP values can help you see which features are most important for the model and how they affect the outcome.

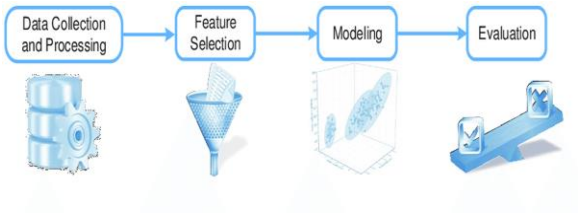

Fig.Pipeline

A) Data Collection and Processing:

In the data collection step for visualizing machine learning through dynamics, initially, relevant datasets are gathered from various sources such as healthcare datasets, customer datasets related to telecom companies, etc. It's essential to consider ethical considerations and implement data privacy measures during this process. Once the data is collected, it undergoes exploratory analysis (EDA) to understand its distribution and characteristics before further preprocessing steps.

The dataset used comes from an Iranian telecom company, with each row representing a customer over a year. Along with a churn label, there is information on the customer's activity, such as call failures and subscription length. This dataset contains a total of 14 attributes.

	Call Failure	Complaints	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS
0	8	0	38	0	4370	71	
1	0	0	39	0	318	5	
2	10	0	37	0	2453	60	
3	10	0	38	0	4198	66	
4	3	0	38	0	2393	58	

5 rows

Fig. Dataset

This data is then preprocessed to clean noise, handle missing values, and standardize features to ensure consistency and quality.

B) Feature Selection:

The collected data typically includes features such as call failure, subscription length, the charge amount, seconds of use, frequency of use, frequency of SMS,

distinctly called numbers, age group, tariff plan, status, age, customer value, and churn. Selecting optimal features that contribute most to model training.

By carefully choosing relevant features, we can reduce dimensionality and highlight the most informative aspects of the data, making it easier to interpret and visualize model behavior.

C) Modeling:

Modeling involves creating graphical representations that illustrate how algorithms learn and adapt over time. These visualizations often depict the interplay between input data, model parameters, and the optimization process. It helps both practitioners and non-experts understand complex machine learning concepts intuitively, fostering insights into algorithm behavior, identifying potential issues, and guiding improvements in model design and training strategies.

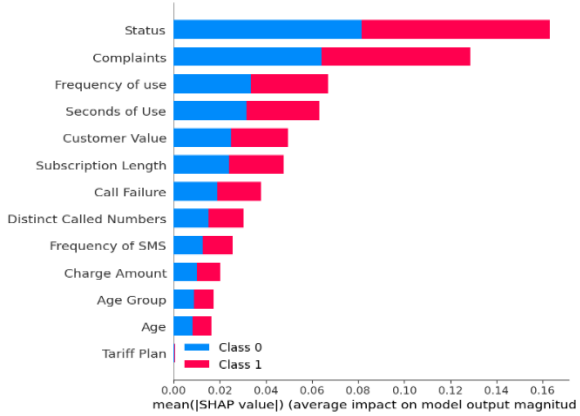

Fig. Modeling

The summary plot shows the feature importance of each feature in the model. The results show that “Status,” “Complaints,” and “Frequency of use” play major roles in determining the results.

D) Evaluation:

Exploring model output in an intuitive and user-friendly manner.

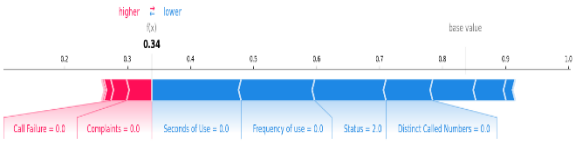

Fig. Evaluation

You can see all of the features with the value and magnitude that have contributed to a loss of customers. It seems that even one unresolved complaint can cost a telecommunications company.

RESULT

Visualizing machine learning through dynamics involves employing various techniques to understand how models evolve during training. One common approach is to plot the loss function over epochs, showcasing the optimization process. As training progresses, the loss typically decreases, reflecting the model's improving performance. Additionally, visualizing decision boundaries in classification tasks provides insights into how the model separates different classes in the feature space. Dynamic visualizations, such as animated plots showing parameter updates or feature space transformations, can offer intuitive explanations of complex machine-learning concepts. These visualizations not only aid in understanding the inner workings of algorithms but also help in diagnosing issues like overfitting or underfitting. Overall, visualizing machine learning dynamics enhances comprehension, facilitates model interpretation, and fosters improvements in algorithm design and training strategies.

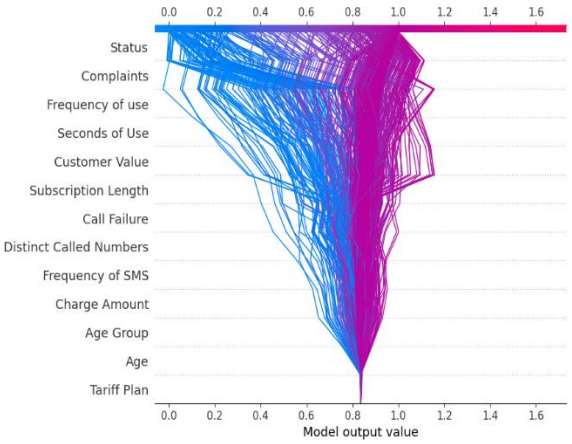

Fig. Output

Each plotted line on the decision plot shows how strongly the individual features contributed to a single model prediction, thus explaining what feature values pushed the prediction.

CONCLUSION

The experiments show that interactively selected training samples can reach higher performance quickly than randomly selected samples. This approach provides a new way to train a machine-learning model using a small set of training samples. since human knowledge and perceptual instincts are used in the

selection of the training samples, this approach is potentially smarter and more efficient than traditional "big data" solutions. It is particularly useful for applications where high-quality "big data" is not readily available or if the collection and labeling of the data is too expensive (e.g. in some biomedical data analysis applications). On the other hand, since this approach requires humans in the learning loop, it may not be suitable for applications that require total automation (e.g. in real-time robot vision).

REFERENCE

[1] Thomas J, Cook K: *Illuminating the Path: Research and Development Agenda for Visual Analytics*. IEEE-Press; (2005)

[2] Amershi S, Cakmak M, Knox WB and Kulesza T, 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4), pp.105–120.

[3] Ware M, Frank E, Holmes G, Hall M and Witten IH, 2001. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3), pp.281–292.

[4] Paiva JG, Florian L, Pedrini H, Telles G, Minghim R, 2011. Improved similarity trees and their application to visual data classification. *IEEE TVCG* 17 (12), 2459–2468

[5] Xia Jing, Chen Wei, Hou Yumeng, Hu Wanqi, Huang Xinxin, Ebert David S. DimScanner: A Relation-based Visual Exploration Approach Towards Data Dimension Inspection. *VAST* 2016.

[6] Liu Shixia, Wang Xiting, Liu Mengchen, Zhu Jun. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics* 1 (2017) 48–56.

[7] Zahavy T, Ben-Zrihem N, Mannor S 2016. Graying the black box: Understanding dqns. In: *ICML* pp. 1899–1908.

[8] Rauber PE, Fadel S, Falcao A, Telea A, 2017. Visualizing the hidden activity of artificial neural networks. *IEEE TVCG* 23 (1), 101–110.

[9] Tzeng FY, Ma KL 2005. Opening the black box - data-driven visualization of neural networks. In: *IEEE Visualization*, pp. 383–390. 10.1109/VISUAL.2005.1532820.

[10] Harley AW, 2015. *An interactive node-link visualization of convolutional neural networks*. In: International Symposium on Visual Computing Springer, pp. 867–877.

[11] Streeter MJ, Ward MO, Alvarez SA, 2001. *Nvis: An interactive visualization tool for neural networks*.

[12] Liu M, Shi J, Li Z, Li C, Zhu JJH, Liu S, 2017. Towards a better analysis of deep convolutional neural networks. *IEEE TVCG* 23 (1), 91–100. http://dx.doi.org/10.

[13] Yosinski Jason, Clune Jeff, Nguyen Anh, Fuchs Thomas, and Lipson Hod. *Understanding Neural Networks Through Deep Visualization*. ICML Workshop on Deep Learning, 2015.

[14] Zintgraf Luisa M, Cohen Taco S, Adel Tameem, Welling Max. *Visualizing Deep Neural Network Decisions: Prediction Difference Analysis*. International Conference on Learning Representations (ICLR) 2017.