Visualizing A Walk Through the Random Forest

Samuel Meyer, Yiyi Chen, and Marti A. Hearst*
School of Information, UC Berkeley

ABSTRACT

Well-designed visualizations have an important role to play to aid in the public's understanding of algorithms. This work presents a set of design principles for using visualization to explain machine learning algorithms specifically, and demonstrates these principles applied to the operations of the random forest algorithm.

Index Terms: H.5.0 [Information Interfaces and Presentation]: General

1 INTRODUCTION

Machine learning is increasingly important in research, business and society today. As machine learning techniques proliferate, the need to explain them to a wider audience becomes ever more pressing. Some recent high-profile efforts to promote the public's understanding of machine learning include the non-profit AI for Good Foundation and the online journal Distill.pub, whose goal is to publish clear explanations of machine learning algorithms. We suggest that information visualization can help explain machine learning algorithms if placed in the right context.

2 BACKGROUND

Algorithm visualization can be classified into static, animated, and interactive designs. Static views include charts and graphs that illustrate the steps of an algorithm, and date back to the flow charts of the earliest days of computer programming. The efficacy of animation in the visualization of algorithms to date is mixed; however, interactive visualizations, which allow user control of disclosure and control of steps of animation, can be beneficial [5].

Hundhausen et al. [3] conducted a meta-analysis that found that "most successful educational uses of AV technology are those in which the technology is used as a vehicle for actively engaging students in the process of learning algorithms" such as performing what-if analysis of algorithm behavior and doing prediction exercises, but not through transferring knowledge via the image.

Most approaches to algorithm animation place the visualization within a monolithic stand-alone view. Instead, we advocate interweaving interactive visualizations within a narrative using well-motivated iconography, layout, and exposition [1, 4].

The starting point for our design is a website that explains decision trees [6]. Via a scrolling HTML5 canvas-style interface, this design evocatively illustrates the major components of machine learning, including an introduction to classification problems, identifying features, separating training and test sets, and an impressive animation of the final predictions of the decision tree in which data points roll down the branches. This design has numerous merits, but some aspects of the animations are problematic from a usability perspective, and there is a missed opportunity in the lack of narrative infographics to illustrate the meaning of the selected dataset.

*{meyer_samuel, yiyi.chen, hearst}@berkeley.edu

Figure 1: Introducing the theme and the example problem.

3 DESIGN CHOICES FOR ML EXPLANATIONS

To better contextualize a visualization of a machine learning explanation, we suggest the following design principles:

(1) Create an interesting and approachable narrative. Our method introduces a playful example used in a consistent and meaningful manner throughout. Because the machine learning technique under study, the random forest, already has an evocative name, we made use of that name as the starting point for the conceptual metaphor. Given the forest metaphor, we choose a mushroom classification problem from the UCI data repository.[1] Mushrooms have easily iconified shapes, and the classification problem of deciding which are poisonous is both interesting and easily understandable, thus making this dataset a good one for exposition.

(2) Use animation solely to aid understanding while always ensuring that all steps are controllable and reversible by the user. Explain the meaning of each animation.

(3) Use visual components (color, iconography, gestalt properties) consistently throughout the design.

4 THE DESIGN

We divided the process of training and using random forest into various sections to guide users. Each has a header on the web page.[2]

4.1 The Mushroom Problem

We begin with an introduction to the problem of deciding whether a mushroom is poisonous or not. We introduce the dataset and some example features to give readers an understanding of the value of classification problems (see Figure 1).

4.2 Decision Trees

Random forests build on decision trees, but because decision trees are visualized elsewhere, we decided to limit our explanation to a basic introduction in text along with an annotated dynamic decision tree for readers to explore. This single decision tree visualization appears again in Figure 3 in a slightly more complex form.

4.3 Fallacy of Individual Predictors

This section introduces the idea of ensemble predictors. Using icons, we illustrate how each decision tree can make mistakes predicting which mushrooms are poisonous, but together they can be better

[1] https://archive.ics.uci.edu/ml/datasets/Mushroom
[2] https://waternova.github.io/random-forest-viz/

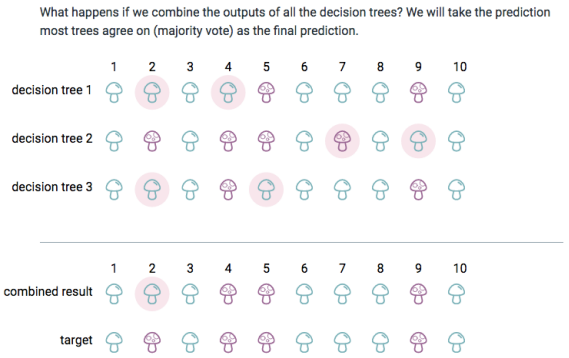

Figure 2: Icons, consistent colors, and gestalt properties of similarity to illustrate ensemble prediction.

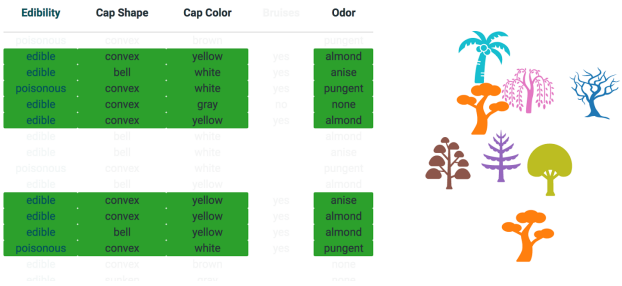

Figure 3: Creation of decision trees from samples of features; animation of the table signals randomness of selection.

predictors (see Figure 2). We use the mushroom theme to maintain consistency throughout the visualization.

4.4 Training of Trees

This section crucially links the concepts of feature and data selection to the creation of each decision tree in the forest. A table shows rows of features for each mushroom. Each time the user selects "Next", a subset of rows and columns of the table are highlighted to indicate the random selections from both to create a sample to train from. Training is indicated with a pulsing message, after which a new tree is animated as emerging from the selected data and growing in the forest on the right (see Figure 3). The differently shaped trees can be inspected in detail in the following visualization.

4.5 Training Results

Each decision tree trained in the previous step can be inspected in this view, which demonstrates that each tree has different decision nodes because it was trained on different data and features (Figure 4). Clicking on a tree icon at the top places the root of the decision tree on the far left in the rectangular region below. Each component can be expanded and collapsed by the user.

4.6 Making a New Prediction

Finally, we show how decision trees in a random forest work together to decide whether a new mushroom is poisonous or not. We animate copies of the mushroom being sent to each tree. When the user clicks "Next", each tree decides whether the mushroom is safe to eat or not. Based on majority vote, the forest decides that the mushroom is safe to eat (see Figure 5). We animate dots for each vote to keep continuity from each tree to the total votes.

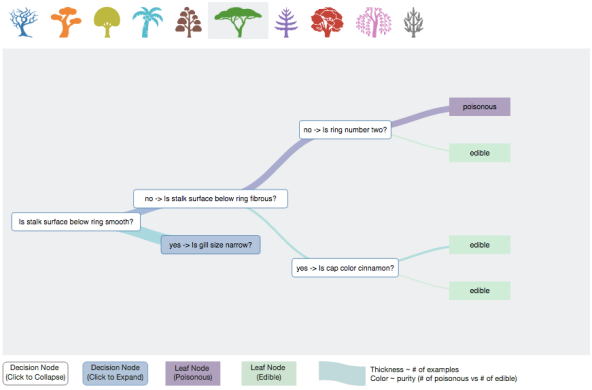

Figure 4: Inspecting component decision trees to see how they differ.

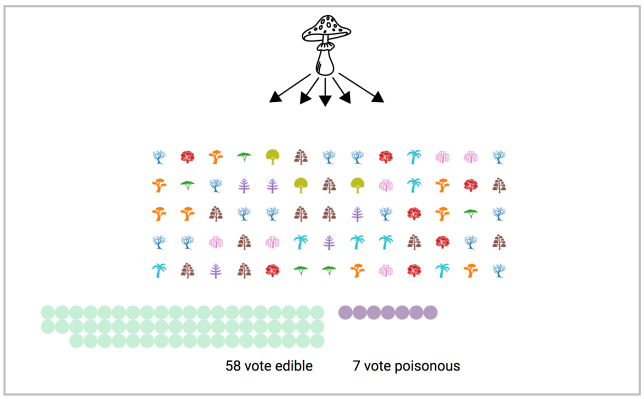

Figure 5: Final animation demonstrates Random Forest voting algorithm. Animation of voting process not shown.

5 Conclusions

We have suggested a set of design principles for explaining machine learning algorithms making use of narrative infographics to retain a consistent theme and user-controlled animations that illustrate the workings clearly. Because machine learning algorithms make use of a dataset and domain problem, they are especially ripe for illustration with iconography, which has been shown to aid memorability if done well [2]. We illustrate these principles with a demonstration on the random forest machine learning algorithm.

References

[1] J. Appleman, A. Gupta, A. Rajagopal, J. Shishido, and M. A. Hearst. Exploring data for fun and profit: Case study of jeopardy! In *Proceedings of IEEE Infoviz, Posters*, 2015.

[2] S. Haroz, R. Kosara, and S. Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *CHI*, 2015.

[3] C. D. Hundhausen, S. A. Douglas, and J. T. Stasko. A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages & Computing*, 13(3):259–290, 2002.

[4] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16:1139–1148, 2010.

[5] J. Urquiza-Fuentes and J. Á. Velázquez-Iturbide. A survey of successful evaluations of program visualization and algorithm animation systems. *TOCE*, 9:9:1–9:21, 2009.

[6] S. Yee and T. Chu. A visual introduction to machine learning. `www.r2d3.us/visual-intro-to-machine-learning-part-1/`. 2017-05-31.