Expose - Explainability Machine Learning - Visualization of Random Forests

Fabio Rougier

April 8, 2022

Contents

1 Introduction

Random Forests (RF) are a powerful ensemble method with a low barrier of entry. Because of their ease of use and performance they are used in many applications. However, they fall short when it comes to transparency. RFs yield a multitude of valuable insights into their decision making and the data they process. Visualizing this is usually a challenge because of the inherent scale that of a RF. In this work we want to look at the approaches trying to overcome this limitation.

2 Visualizing Decision Trees

An obvious approach to visualizing a RF is to inspect the decision trees, constituting the RF. One example for the visualization of a decision tree is *BaobabView* (Van Den Elzen & Van Wijk, 2011). As many other approaches visualizing decision trees, it utilizes Node-Link Diagrams (NLDs) as its' main visualization. A confusion matrix yields additional insights into the relations of the underlying features. However this visualization does fall short when trees grow too large, as it becomes hard to inspect every individual node and branch of the tree. This is one of the problems that *TaxonTree* tries to overcome (Parr, Lee, Campbell, & Bederson, 2003). It uses a tree visualization approach that is scalable for large trees by adding the possibility to zoom, browse and search the tree. While this does deal help if a tree grows too large, it does not provide any insight on the general structure of the tree as a whole.

Generalizing either of the approaches towards RFs is also non trivial. Both simply lack the scalability in the desired dimension. It also leads into a dangerous of focusing too much on the structure of individual trees. RFs - as all ensemble methods - arrive at their decisions by combining the decisions of the individual trees. Inspecting and even understanding individual trees, will only yield limited insights over the RF.

3 Visualizing Random Forests

To fully understand the structure and decision making of a RF, many aspects of the RF have to be conveyed by the visualization. First of all the typical indicators like a *mean squared error* or a *mean average error* of any machine Learning model give an indication of the overall performance. Adiitionaly there are some metrics specific to RFs that should be included to get a deeper understanding of the specific instance of the RF, like the *mean impurity* of the individual trees or the *out of bag error*. With its' unique way of giving a distance measure for features, a confusion matrix can also yield valuable insights to the relations of the features and how the RF interprets them. As stated before, the visualization of RFs heavily relies on how it overcomes the scalability issues of RFs.

3.1 RAFT

The RAFT visualization is one of the first of its' kind and features all of the aforementioned criteria.

4 Summary

(Zhao, Wu, Lee, & Cui, 2018)

References

Parr, C. S., Lee, B., Campbell, D., & Bederson, B. B. (2003). *Taxontree: Visualizing biodiversity information* (Tech. Rep.). MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES.

Van Den Elzen, S., & Van Wijk, J. J. (2011). Baobabview: Interactive construction and analysis of decision trees. In *2011 ieee conference on visual analytics science and technology (vast)* (pp. 151–160).

Zhao, X., Wu, Y., Lee, D. L., & Cui, W. (2018). iforest: Interpreting random forests via visual analytics. *IEEE transactions on visualization and computer graphics*, *25*(1), 407–416.