treeheatr: an R package for interpretable decision tree visualizations

Authors

- **Trang T. Le**
  ⓘ 0000-0003-3737-6565 · ☉ trang1618 · 🐦 trang1618
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104

- **Jason H. Moore**
  ⓘ 0000-0002-5015-1099 · ☉ EpistasisLab · 🐦 moorejh
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104 · Funded by National Institutes of Health Grant Nos. [LM010098, LM012601, AI116794]

Abstract

Summary

treeheatr is an R package for creating interpretable decision tree visualizations with the data represented as heatmaps at the tree's terminal nodes. Going beyond the if-then step-by-step logic of a decision tree, the inclusion of a heatmap offers a broader view of the classification or regression problem and provides meaningful clarification of different node splits in the tree. Working harmoniously with other packages, *treeheatr* empowers the user with refined controls over the statistical threshold and presentation of the tree and heatmap.

Availability and implementation

The *treeheatr* package is freely available under the permissive MIT license at https://trang1618.github.io/treeheatr. It comes with a detailed vignette that is automatically built with continuous integration.

Contact

ttle@pennmedicine.upenn.edu

Supplementary information

Introduction

Tree-based algorithms such as random forests and gradient boosted trees are widely used techniques that comprise an important section of supervised machine learning. Decision trees are primitive constituents of many important including . Visualizing and intepreting their building blocks, the single decision trees, are the first steps toward understanding these more complex tree-based structures. Nonetheless, existing visualization softwares frequently treat all nodes similarly, leaving limited options for improving visualization at the leaf nodes. Currently, at leaf nodes, state-of-the-art libraries such as Python's dtreeviz draw pie charts.

In a visualization, it is difficult to find the sweet spot on the spectrum from enhancing understanding to overloading information.

an opportunity for the data to be shown at the leaf nodes as a heatmap.

heatmap: complimentary of histogram at non-leaf nodes. correlation: implied from clustering

Materials and methods

treeheatr: a simple example

This example visualizes the conditional inference tree model built to predict whether or not a patient has diabetes from a dataset provided by the National Institute of Diabetes and Digestive and Kidney Diseases [1]. This dataset of female patients at least 21 years old of Pima Indian heritage near Phoenix, Arizona was downloaded from Kaggle and has eight features: age, number of times pregnant, plasma glucose concentration, diastolic blood pressure, skin fold thickness, 2-hour serum insulin, body mass index and diabetes pedigree function. Detailed descriptions of these variables and data source can be found on the Kaggle page.

The following lines of code computes and visualizes the conditional decision tree along with the heatmap containing features that are important for building this model (Fig. 1):

```
heat_tree(
  data = diabetes,
  target_lab = 'Diabetes status',
  label_map = c(`0` = 'Negative', `1` = 'Positive')
)
```

The heat_tree() function takes a data frame, a character string indicating the column name associated with the outcome/phenotype (e.g., Diabetes status) and other optional arguments such as the mapping of the outcome label.

Figure 1: example/diabetes.png

Conclusion

In this paper, we presented the new type of integrated visualization of decision trees and heatmaps, which provides a comprehensive data overview as well as model interpretation. We demonstrated that [...] The visualization is implemented in an easily installed package with a detailed vignette. Released as open source software, {treeheatr} hopes to receive contribution from other developers.

Future works on *treeheatr* include enhancements such as supporting heatmap visualization of a holdout set and highlighting the tree branches that point to a specific holdout sample.

Acknowledgements

References

1. **Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus**
   Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, R. S. Johannes
   Proceedings of the Annual Symposium on Computer Application in Medical Care (1988-11-09)
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/
   PMCID: PMC2245318

2. **Ggplot2: elegant graphics for data analysis**
   Hadley Wickham
   Springer (2009)
   ISBN: 9780387981406

3. **partykit: A Modular Toolkit for Recursive Partytioning in R**
   Torsten Hothorn, Achim Zeileis
   Journal of Machine Learning Research (2015) http://jmlr.org/papers/v16/hothorn15a.html

4. **martin-borkovec/ggparty**
   Martin Borkovec
   (2020-04-25) https://github.com/martin-borkovec/ggparty