Exploring the Impact of Decision Tree Depth

Alic Szecsei
University of Iowa
alic-szecsei@uiowa.edu

Diego Castaneda
University of Iowa
diego-castaneda@uiowa.edu

Willem DeJong
University of Iowa
willem-dejong@uiowa.edu

ABSTRACT

Lobortis scelerisque fermentum dui faucibus in ornare quam viverra orci sagittis eu volutpat odio facilisis mauris sit amet massa vitae tortor condimentum lacinia quis vel eros donec ac odio tempor orci dapibus ultrices in iaculis nunc sed augue lacus viverra vitae congue eu consequat ac felis donec et odio pellentesque diam volutpat commodo sed egestas egestas fringilla phasellus faucibus scelerisque eleifend donec pretium vulputate sapien nec sagittis aliquam malesuada bibendum arcu vitae elementum curabitur vitae nunc sed velit dignissim sodales ut eu sem integer vitae justo eget magna fermentum iaculis eu non diam phasellus vestibulum lorem sed risus ultricies tristique nulla aliquet enim tortor at auctor urna nunc id cursus metus aliquam eleifend mi in nulla posuere sollicitudin aliquam ultrices sagittis orci a scelerisque purus semper eget duis at tellus at urna condimentum mattis pellentesque id nibh tortor id aliquet lectus proin nibh nisl condimentum id venenatis a condimentum vitae sapien pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas sed tempus urna et pharetra pharetra massa massa ultricies mi quis hendrerit dolor magna eget est lorem ipsum dolor sit amet consectetur adipiscing elit pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas integer eget aliquet nibh praesent tristique magna sit amet purus gravida quis blandit turpis cursus in hac habitasse platea dictumst quisque sagittis purus sit amet volutpat consequat mauris nunc congue nisi vitae suscipit tellus mauris a diam maecenas sed enim ut sem viverra aliquet eget sit amet tellus cras.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; *Supervised learning by classification*; *Classification and regression trees*; Cross-validation.

KEYWORDS

decision trees, model selection

1 BACKGROUND AND MOTIVATION

Machine learning algorithms fall into one of two categories: classification and regression. In regression, input data is consumed and transformed by a function which can produce values along a numeric range; in classification, data is categorized into a finite number of possible values, and machine learning algorithms seek only to label data sets. These classification problems can use both numeric and categorical data attributes. For purely categorical data, decision trees are a way to build machine learning algorithms that are understandable by both humans and machines. Rather than a sort of "black box" in which numbers are passed in and answers are returned, there is a logical tree hierarchy that is familiar to anyone who has played the game Twenty Questions.

Decision trees require both a way to construct them, and a way to use them to categorize data. This categorization is a trivial tree traversal, and so many of the innovations regarding decision trees has to do with their construction. Most decision tree construction algorithms use a two-phase approach: first a *growing* phase, followed by a *pruning* phase. In the growing phase, the decision tree is built out, trying to fit the provided training data as closely as possible. To combat over-fitting, the pruning phase determines which branches of the tree are too "noisy" using χ^2 tests and removes them.

Additionally, decision tree models can be constrained by size to combat overfitting. Russell and Norvig [3] showcase an implementation of restricting a decision tree to be beneath a maximum size by generating the tree in breadth-first fashion, and stopping when the maximum number of nodes has been reached. As stated in Garofalakis, Hyun, Rastogi, and Shim [1], there is no point in creating a branch when it is guaranteed to be pruned later.

The amount of pruning which occurs is heavily dependent on the data set chosen. As such, the exact values which optimize our trained decision trees are relatively unimportant. However, our goal was to examine how impactful depth-based pruning was, and how much it assisted with reducing overfitting.

2 METHODS

The foundation of our decision tree algorithm was the one provided by Russell and Norvig [3] which, in turn, is based on the ID3 algorithm [2].

We selected four data sets from the Machine Learning Repository at http://archive.ics.uci.edu/ml/datasets.php. We then ran each through the DecisionTreeLearning algorithm with maximum depths ranging from 1 to 10 and performed k-fold cross-validation to determine error, with $k = 4$.

3 RESULTS

4 DISCUSSION

ACKNOWLEDGMENTS

This work was supported in part by the University of Iowa.

REFERENCES

[1] Minos Garofalakis, Dongjoon Hyun, Rajeev Rastogi, and Kyuseok Shim. 2000. Efficient Algorithms for Constructing Decision Trees with Constraints. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*. ACM, New York, NY, USA, 335–339. https://doi.org/10.1145/347090.347163

[2] J. R. Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (01 Mar 1986), 81–106. https://doi.org/10.1007/BF00116251

[3] Stuart J. Russell and Peter Norvig. 2010. *Artificial intelligence: a modern approach* (3rd ed.). Prentice Hall.

A RESEARCH METHODS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi malesuada, quam in pulvinar varius, metus nunc fermentum urna,

Function DecisionTreeLearning(*examples*, *attributes*, *parent_examples*, *depth*)

```
if examples is empty then
    return Plurality-Value(parent_examples)
else if depth = 0 then
    return Plurality-Value(examples)
else if all examples have the same classification c then
    return a leaf node c
else if attributes is empty then
    return Plurality-Value(examples)
end
A ← argmax_{a∈attributes} Importance(a, examples);
tree ← a new decision tree with root test A;
foreach value v_k of A do
    exs ← {e : e ∈ examples and e.A = v_k};
    subtree ←
     DecisionTreeLearning(exs, attributes - A, examples, depth - 1);

    add a branch to tree with label (A = v_k) and subtree
     subtree;
end
return tree
```

id sollicitudin purus odio sit amet enim. Aliquam ullamcorper eu ipsum vel mollis. Curabitur quis dictum nisl. Phasellus vel semper risus, et lacinia dolor. Integer ultricies commodo sem nec semper.

B ONLINE RESOURCES

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi malesuada, quam in pulvinar varius, metus nunc fermentum urna, id sollicitudin purus odio sit amet enim. Aliquam ullamcorper eu ipsum vel mollis. Curabitur quis dictum nisl. Phasellus vel semper risus, et lacinia dolor. Integer ultricies commodo sem nec semper.