GENERALIZED DECISION TREE

A PREPRINT

Jarred Parr
Department of Computer Science
Grand Valley State University
parrjar@mail.gvsu.edu

March 12, 2019

ABSTRACT

Decision trees are an extremely robust data structure which follows a more logical design pattern with reproducible output. Boasting robustness to outliers and practically baked-in support for forest regression, decision trees have become the statistical tool of choice for data scientists looking for rock-solid classification without the overhead of a full fledged neural network. This paper explores an attempted hyper-optimization of such a system via C++.

1 Introduction

This projected attempted an ambitious goal. Not only was a successful attempt made to construct the decision tree algorithm with barebones, modern C++ (no libraries at all), but a further attempt was made at parallelism of a forest regressor implementation on the origin decision tree idea. Random forests were used because of their ability to squeeze every bit of optimization out of an algorithm. Multiple runs can garner different results and when taken further with techniques like gradient boosting, you can have a very powerful model without much extra overhead. This project accomplished two of the three goals here with differing levels of success therein. The decision tree in its most basic form worked fine without much extra work needed. However, when implementing parallel random forests, there was a bit more difficulty when attempting to apply this same algorithm. This was primarily due to memory issues when large numbers of trees were added and, in quite a number of cases, inaccurate results.

2 Algorithm

Of the two available algorithms, ID3 was used in favor of the other popular options. It seemed to have the best reviews when compared to other approaches, and it also was the easiest to understand. The ID3 Algorithm was used as defined below:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

$$SplitInfo_A(D) = -\sum j = 1^v \frac{|Dj|}{|D|} log_2(\frac{|D_j|}{|D|})$$

$$Gain(A) = info(D) - info_A(D)$$

$$info_A(D) = \sum j = 1^v \frac{|Dj|}{|D|} info(D_j)$$

$$info(D) = -\sum i = 1^m pilog2(pi)$$

$$p_i = \frac{|C_{i,D}|}{|D|}$$

The ID3 algorithm was the most straightforward algorithm to use because of its ease of implementation. Even though recursion tends to absorb a bit more of the overall memory space, when handled in an iterative manner (when useful) it was able to have its overall footprint shrunk significantly when used on very large data sets. Of the data used, the provided sets were used to determine if the algorithm worked well, and from there, the Microsoft Malware Database hosted on Kaggle.com was used to take things a step further. Boasting about 3.8gigs of total data, it presented an immeasurable amount of difficulty working with this data set because of how much ram not only it used, but also the algorithm when it began computation. This was used as a benchmark to determine the overall speed at which the algorithm could perform in a real life scenario.

2.1 Performance

The algorithm boasted shockingly fast performance for what was considered at the time a naive implementation. When using the baseline decision tree, most data sets, even the largest of the provided, took less than one second to complete. However, when using the Microsoft Data set, it took an upwards of 6 hours to finally see an interpretable output from the run. To massage the random forest layer, the data needed to be handled chunk-by-chunk in memory to keep the system from locking up and having the process killed by the OS. A chunk algorithm was devised to manage this runtime, but the algorithm trained about 50% slower taking about 9 and a half hours to finally produce something, which was not very accurate in the end. It was determined that this was partly with how the data was chunked as the previous information had to be cleared from memory for the new, large segment to be loaded in. Keeping track of locations in the file and manually managing memory put a wrench in the process overall.

3 Random Forest

The Random Forest was a bit more of a challenge. As a whole, it was a bit more difficult to get things running in parallel due to its complexity. A toy version of iteratively running the trees and averaging their results was experimented with, but in the end, it ended up taking an astoundingly long time for only minuscule gains in overall accuracy. As a result, a parallel approach was explored. Unfortunately, this ended up being only minimally better and extremely unstable. The code has been omitted because of this. There were problems with race conditions inside of the recursive sections, and as the number of threads began to increase to ever larger numbers, the algorithm began to absorb system resources so quickly that the entire UI locked until the OS reaped the process. It is clear that a serial mindset when applied to an algorithm like this may not be the best in the long run. Algorithms such as this have a clear need to be designed with parallelism from the start. OpenMP directives can only do so much for your algorithm until a complete rewrite is needed. It was a very interesting system to try, though.

4 Growth Areas

Many pieces of previous projects have been about how to use the modern C++ tools to accomplish some of the more difficult tasks presented, however, it is clear that, in this case, there was a lot of chance to grow in the design of parallel algorithms. Parallelism has recently been a hot topic amongst machine learning research groups for its ability to get models working faster and faster when applied to high-volume, high-throughput datasets that are commonly seen in a large enterprise environment. As a result, learning how these techniques can be applied when implementing from scratch allows the implementer to take things to a deeper level and expose a higher level of expertise than previously thought. When working through this, it was clear that there were still some things that I need to refine about my process when looking toward adding parallelism to upcoming projects and side work.

5 Results

The results for the algorithm were quite good. When compared to the built-in algorithms of the scikit-learn library, there was a very clear marker that the implementation of this project was at least reasonable. In almost every situation the C++ implementation came within 5% of the scikit-learn run. The examples are compared in the following table.

As can be seen, the runtimes were very close except when exposed to a massive dataset. The Scikit-Learn was run on the same system as the custom implementation and the custom implementation is the clear winner. This is to be expected as python has a significantly slower performance when compared to optimized C++.

6 Conclusion

Overall, this project presented the largest challenge of any algorithm that has been attempted thus far in the class. Using C++ certainly doesn't help make things easy in this regard, but understanding the algorithm and overcoming

Table 1: Dataset Runtime Comparison

Name	Custom	Scikit-Learn
In-Class	< 1 Second	< 1 Second
Contact Lenses	< 1 Second	< 1 Second
Cars	< 1 Second	1.1 Seconds
Microsoft	6 Hours	9 Hours

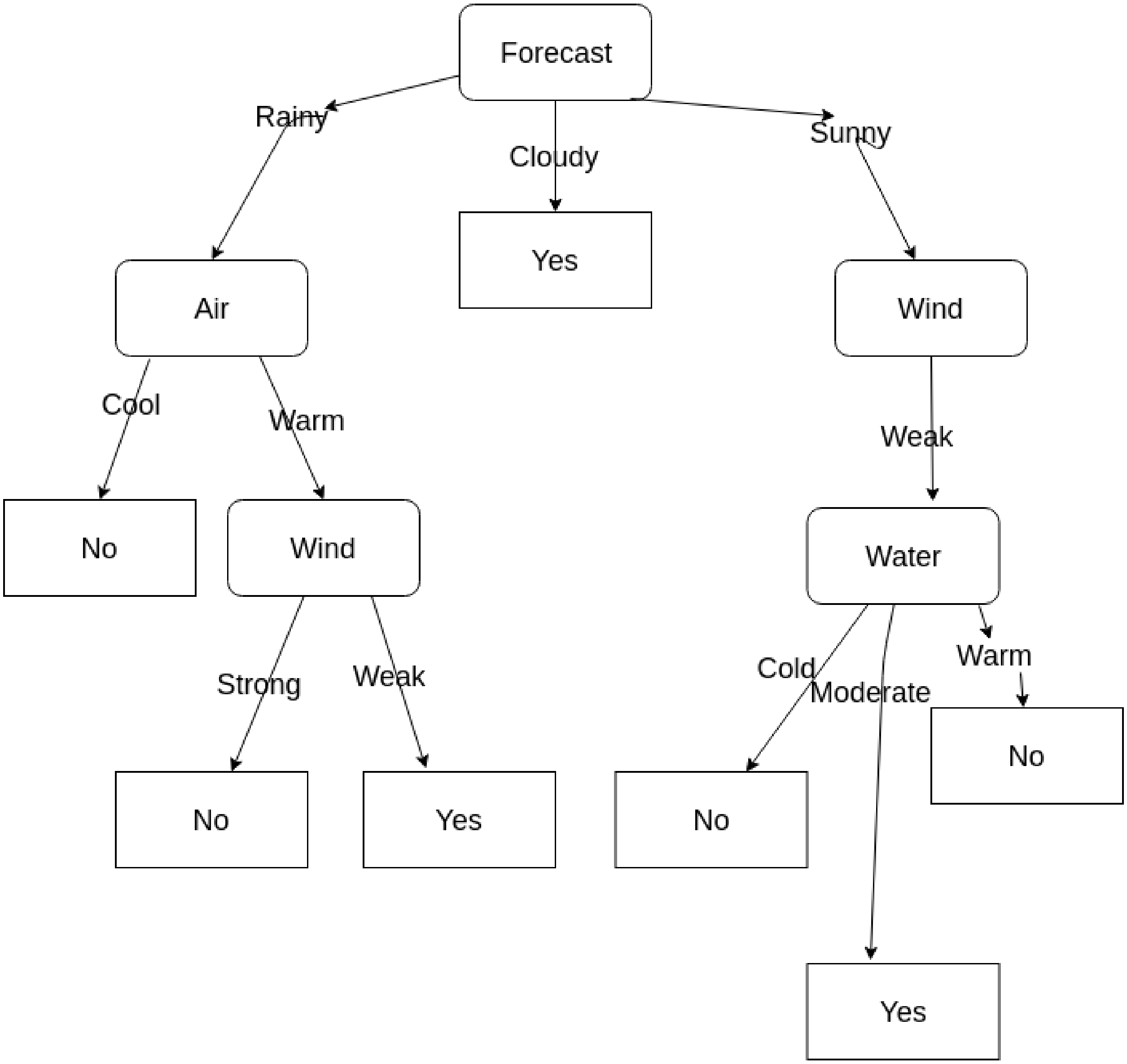

Figure 1: The Test Data Graph

bottlenecks imposed by the arguably dense recursion definitely took some time. Overall this project was by far the most interesting and useful of the algorithms explored so far. In the future I plan to spend more time learning the inner workings of the algorithm to prevent losing time to silly errors and misunderstandings in the future. Also, the goal is to get a GPU-accelerated project working at some point in the semester.