A Review and Comparison of Boosting Methods

Alexander Oleinik
alxndr@bu.edu

Tayler Pauls
tayler@bu.edu

Shen Shen
shs2016f@bu.edu

Abstract

Our team is performing a thorough review of published boosting methods. Our final presentation and report will detail the differences in methods used in meta-algorithms such as AdaBoost and Gradient Boosting. To outline the impact of these differences, we will perform empirical comparisons on multiple datasets to demonstrate their effect on performance. Additionally, we will apply boosting to a real-world application.

1. Introduction

Boosting algorithms are a family of machine-learning ensemble approaches that turn weak classifiers into strong classifiers; they improve classifiers that only outperform random predictions by a small amount. In 1990, Robert Schapire formulated a rigorous explanation of why boosting is possible, He and Yoav Freund pioneered the first practical boosting algorithm - AdaBoost. Although AdaBoost remains popular, there has been significant innovation in the field of Boosting since 1997. Our project will dissect the theoretical differences among Boosting algorithms and provide empirical results to depict data-set categories where particular Boosting methods excel or perform well.

1.1. Methods

In order to efficiently compare Boosting methods, we will develop an environment to automatically apply a range of boosting algorithms to a dataset and report performance metrics for each method. Depending on availability, the boosting implementations will be sourced from established codebases otherwise they will be built manually. We will track boosting performance while adjusting feature parameters such as artificial noise and separability. Then we will compare the effects of various black art adjustments on the data and monitor correlations between black art parameters and classification performance. Additionally, we will perform classification with a control classification tree model as a baseline in all our tests.

1.2. Data Sets

To start, we will generate data sets through simulations to test our boosting framework. This way we will be able to control the separability of the data sets and the number of classes. This will allow us to observe the dependence of boosting performance on the generation parameters compared to the classification tree performance.

1.3. Goals

- Convey the theoretical framework of boosting and its significance with the context of the topics previously covered in class. Additionally, highlight the motivation and advantages presented by the boosting class of meta-algorithms.

- Have a demo which will generate expository data sets and perform classification via Boosting and classification trees. Our demonstration should depict key Boosting concepts such as adaptivity and decision trees.

- Present the correlation between parameters of the black art transformations and the performance of Boosting methods compared to classification trees. Visualisation of the mechanics of AdaBoost and xgBoost.

  Note: Performance metrics will be measured with CCR and Recall.

1.3.1 Fallback

If we are unable to formulate expository results with a real dataset, we will focus our time on generating smaller interesting datasets and exploring boosting methods other than AdaBoost and xgBoost. Our main goal is to provide some theory and intuition about the Boosting family of algorithms rather than to apply it to a specific dataset.

1.4. Responsibilities

Our project features literature/theory review programming, and presentation components.

- Theory:
  - ADABoost
  - Gradient Boost
  - Comparison of Boosting Methods
- Implementation:
  - ADABoost demonstration(simulated data set)
  - Gradient Boost demonstration(simulated data set)
  - Real-world application(facial detection)
- Tayler will focus on AdaBoost theory and application to a simulated sample
- Shen will focus on Gradient Boost theory and application to a simulated sample
- Alex will compare boosting methods and apply them to real world samples.

1.5. References

Our project features literature/theory review programming, and presentation components.

1. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016.
2. Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT press, 2012.
3. Schapire, Robert E., and Yoav Freund. Boosting: Foundations and algorithms. MIT press, 2012.
4. McDonald, Ross, David Hand, and Idris Eckley. "An empirical comparison of three boosting algorithms on real data sets with artificial class noise." Multiple Classifier Systems (2003): 161-161.