PADERBORN UNIVERSITY
The University for the Information Society

Department of Electrical Engineering,
Computer Science and Mathematics
Warburger Straße 100
33098 Paderborn

Intelligent Systems Group (ISG)

Seminar paper

Speeding Up Classification Algorithms in Big Data Environments: A review

Clemens Damke

November 8, 2018

Clemens Damke
Speeding Up Classification Algorithms in Big Data Environments: A review
Seminar paper, November 8, 2018
Supervisor: Marcel Wever

Intelligent Systems Group (ISG)
Department of Computer Science
Pohlweg 51
33098 Paderborn

Abstract

TODO

1 Introduction

Over the last few years Big Data processing has become increasingly important in many domains. This increase in the data volume also poses new challenges for machine learning applications. The training time of learners is usually polynomially dependent on the size of the training dataset $\mathcal{D}_{train}$, i. e. $\Omega(|\mathcal{D}_{train}|^k), k \geq 1$. Since training has to be repeated for every iteration of validation and hyperparameter search, always using the entire dataset quickly becomes infeasible. This paper gives an overview of approaches to tackle this problem.

The process of finding an optimal model can in general be split into two phases:

1. **Hyperparameter search:** Finding a vector λ in the hyperparameter space Λ_L of the learner L representing a hypothesis space $\mathcal{H}_\lambda$. A naïve approach for this is a simple grid or a random search over Λ_L. To evaluate the quality of a given λ a parameter search is usually performed which yields a hypothesis $\hat{h} \in \mathcal{H}_\lambda$ that is evaluated using a validation dataset $\mathcal{D}_{valid}$.

2. **Parameter search:** Finding a vector w in the parameter space $W_{\mathcal{H}_\lambda}$, describing a hypothesis $h_w \in \mathcal{H}_\lambda$ given a hyperparameter configuration λ. The goal is to find a hypothesis h_w that minimizes the empirical error on a given test dataset $\mathcal{D}_{test}$. Depending on the learner L, various kinds of optimization methods can be used to find such an hypothesis, e. g. gradient descent or quadratic programming.

This paper is structured according to those phases. Section 2 describes ways to speed up the hyperparameter search. Section 3 then describes how to improve existing optimization methods for parameter search. Most of the techniques described in this paper improve upon orthogonal components of the model finding process which allows combining them.

2 Hyperparameter optimization

[TODO]

3 Parameter optimization

[TODO]

4 Conclusion

[TODO]