INSTITUT FÜR INFORMATIK
Computer Vision, Computer Graphics
and Pattern Recognition

Universitätsstr. 1 D–40225 Düsseldorf

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Classification of data from the ATLAS experiments

Michael Janschek

Bachelor Thesis

Beginn der Arbeit:	10. Dezember 2015
Abgabe der Arbeit:	10. MÃd'rz 2016
Gutachter:	Prof. Dr. Stefan Harmeling
	Prof. Dr. Stefan Conrad

Erklärung

Hiermit versichere ich, dass ich diese Bachelor Thesis selbstständig verfasst habe. Ich habe dazu keine anderen als die angegebenen Quellen und Hilfsmittel verwendet. Alle Stellen, die aus den Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Düsseldorf, den 10. MÃd'rz 2016

Michael Janschek

Abstract

Hier kommt eine ca. einseitige Zusammenfassung der Arbeit rein.

Contents

1 Introduction

This chapter first presents the Higgs Boson Machine Learning Challenge and explains its motivation and goals. It is concluded by an overview of the thesis structure.

1.1 The Higgs Boson and the ATLAS experiments

2013 the Nobel prize in physics acknowledged the discovery of the Higgs Boson. A particle the physicist Peter Higgs predicted to exist, giving mass to other elementary particles.

1.2 The Higgs Boson Machine Learning Challenge

Kaggle is an internet community of data scientists, it hosts several competitions posed by businesses or organizations. Further services involve open datasets, a "Jobs Board" and "Kaggle Rankings", a scoreboard based on performances of community-members in Kaggles competitions.

1.2.1 Motivation

1.2.2 Goal

> The goal of the Higgs Boson Machine Learning Challenge is to explore the potential of advanced machine learning methods to improve the discovery significance of the experiment. No knowledge of particle physics is required. Using simulated data with features characterizing events detected by ATLAS, your task is to classify events into "tau tau decay of a Higgs boson" versus "background." [hig14a]

1.3 Overview

In Chap. 2, we will describe the structure of the challenges dataset [hig14b] and use simple data-analysis methods, to gain first insight about useful features. We will understand the evaluation metric AMS and related formulas we can use as objective functions for optimizing our classifiers. This will be concluded by deriving the formal problem from the challenges task.

Chap. 3 will use first knowledge about the data to choose simple approaches for classification. After these we will describe several more specific and complex methods.

In Chap. 4 we will use the discussed methods and observe their performance on the challenges data. The thesis closes with a discussion about the approaches and their possible influence on other HEP-applications.

2 Understanding the Challenge

blablabla . . . As we continue with Sections 2.4 and 2.5, we will rely on [ABCG+15], as this paper describes these points from the challenge-creators point of view.

2.1 The Data

2.2 notes

Using simulated data with features characterizing events detected by ATLAS, your task is to classify events into "tau tau decay of a Higgs boson" versus "background."

2.3 main

All data provided by the challenge- and the opencern-dataset [hig14b] was created by the official ATLAS full detector simulator in a two-part-process. The simulator first reproduces proton-proton collisions. Then it tracks these via a virtual model of the ATLAS-detector, the resulting data emulates the statistical properties of the real events. Signal-events are generated by tau tau decay of the Higgs Boson, background-events originate from three known processes, which produce radiation similar to the signal. [ABCG+15]

One might expect decent physics-knowledge as key in succeeding in the challenge, the top-participants did not use a lot domain-knowledge for feature- or method-selection. One goal of the challenges organization was to set a task for datascientists without any physics-background.[ABCG+15]

The features $Weight$ and $Label$ were originally only provided in the training-dataset. The data used in this thesis is expanded by complete $Weight$-, $Label$-features and the Kaggle-specific features $KaggleSet$ and $KaggleWeight$.

All features of the dataset are described in Appendix A.

2.3.1 Data analysis

- mathematical
  - weights?
- scatterplots
- histograms

2.4 The formal problem

We use the formal description of the challenge as it is formulated in [ABCG+15].
Let $D = (x_i, y_i, w_i),\ i\ \epsilon\ [1, n]$ be the training sample with n events, where:

- $x_i \epsilon \mathbb{R}^d$ is a d-dimensional vector
- $y_i \epsilon$ {b,s} is the label
- and $w_i \epsilon \mathbb{R}^+$ is a non-negative weight.

The sums of signal and background-signals

$$S = \sum_{y_i=s} w_i$$

and

$$B = \sum_{y_i=b} w_i$$

represent the expected total number of signal and background events.

Let a function $g : \mathbb{R}^d \to \{b, s\}$ be a binary classifier, a set $G_s = \{x : g(x) = s\}$ is called the *selection region*. Our task is to create $g(x_i)$ while maximizing the Approximate Median Significance (AMS). We shall submit an *index set* $\hat{G}_s = \{i : g(x_i) = s\}$ of points that g classifies as signal and an *index set* $\hat{G}_b = \{i : g(x_i) \neq s\}$ while $\hat{G}_s \cap \hat{G}_b = \emptyset$.

2.5 The evaluation

The evaluation of a single submission to the challenge is related to the common practice in particle physics to rate a discovery by its statistical significance, in this case

$$Z = \sqrt{2\left(n \ln\left(\frac{n}{\mu_b}\right) - n + \mu_b\right)} \tag{1}$$

where n is the total number of observed events and μ_b is the expected number of background-events.
Often in particle physics a significance of at least Z=5 (a five-sigma effect) is regarded as sufficient to claim a discovery [ABCG+15].

By estimating $n = s + b$ and $mu_b = b$ in Eq. (4), we get the $Approximate\ Median\ Significance$ (AMS)

$$AMS = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)} \tag{2}$$

which is used by high-energy physicists for optimizing the selection region for stronger discovery significance [ABCG+15].

For the challenge, a regularization-term b_{reg} was introduced as an artificial shift to b to decrease variance of the AMS, as this makes it easier to compare the participants if the optimal signal region was small. "The value $b_{reg} = 10$ was determined using preliminary experiments." [ABCG+15]

This addition to Eq. (2) makes the final evaluation-formula complete:

$$AMS_2 = \sqrt{2\left((s+b+b_{reg})\ln\left(1+\frac{s}{b+b_{reg}}\right)-s\right)} \tag{3}$$

For simplicity, we will call it just AMS, as Eq. (2) will not have further appearances in this thesis.

2.5.1 The Leaderboards

2.5.2 Alternative objective functions

For classification, a data scientist wants to train a classifier on an *objective function*. Properties of the AMS (like using the logarithm) make it difficult to use it as objective function, some alternatives were proposed by the challenge-creators [ABCG+15] and some challenge-participants via the Kaggle-Forum [**?**]

- alternative objective function: $\frac{s}{\sqrt{b}}$
  - only valid when $s \ll b \; and \; b \gg 1$

$$Z = \sqrt{q_0} = \sqrt{2\left(n\ln\left(\frac{n}{\mu_b}\right)-n+\mu_b\right)} \tag{4}$$

$$\frac{s}{\sqrt{b}} \tag{5}$$

(3) (5) (4)

3 Methods of classification

3.1 Logistic regression

- optimizing objective function
- AMS

3.2 k-nn classification

3.3 The winning methods

3.3.1 Neural networks

3.3.2 Regularized greedy forest

3.3.3 XGBoost

4 Results on the Kaggle data

4.1 k-nn classification

4.2 Comparision of all methods

5 Discussion

5.1 Impact of the Challenge

- Ersetzen von AMS durch weighted AUC (papers)
-

[DMNV14] [CH14] [CGG+14]

References

[ABCG+15] ADAM-BOURDARIOS, Claire ; COWAN, Glen ; GERMAIN, Cécile ; GUYON, Isabelle ; KÉGL, Balázs ; ROUSSEAU, David: *Learning to discover: the Higgs boson machine learning challenge.* http://www.http://opendata.cern.ch/record/329, January 2015. – Version 2.3

[CGG+14] COWAN, Glen (Hrsg.) ; GERMAIN, Cécile (Hrsg.) ; GUYON, Isabelle (Hrsg.) ; KÉGL, Balàzs (Hrsg.) ; ROUSSEAU, David (Hrsg.) ; CERN, CRNS (Veranst.): *NIPS 2014 Workshop on High-energy Physics and Machine Learning*. 2014

[CH14] CHEN, Tianqi ; HE, Tong: Higgs Boson Discovery with Boosted Trees. In: *NIPS 2014 Workshop on High-energy Physics and Machine Learning* Bd. 42, 2014 (JMLR: Workshop and Conference Proceedings), S. 69–80. – http://jmlr.csail.mit.edu/proceedings/papers/v42/diaz14.pdf , Accessed: 2016-01-18

[DMNV14] DÍAZ-MORALES, Roberto ; NAVIA-VÁZQUEZ Ángel: Optimization of AMS using Weighted AUC optimized models. In: *NIPS 2014 Workshop on High-energy Physics and Machine Learning* Bd. 42, 2014 (JMLR: Workshop and Conference Proceedings), S. 109–127. – http://jmlr.csail.mit.edu/proceedings/papers/v42/diaz14.pdf , Accessed: 2016-01-18

[hig14a] *Higgs Boson Machine Learning Challenge*. https://www.kaggle.com/c/higgs-boson, 2014. – Accessed: 2016-01-03

[hig14b] *Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014*. http://www.http://opendata.cern.ch/record/328, 2014. – Accessed: 2015-12-10

List of Figures

List of Tables