Incorporation of systematic uncertainties in the training of multivariate methods

Thomas Alef, Eckhard von Toerne

Physikalisches Institut, University of Bonn, Nussallee 12, 53115 Bonn

E-mail: talef@physik.uni-bonn.de, evt@physik.uni-bonn.de

Abstract. Due to the large size of datasets accumulated at the LHC, analysis results are often limited by systematic effects. The application of multivariate analysis techniques such as Boosted Decision Trees (BDTs) or artificial neural nets typically maximises the statistical significance of the results while ignoring systematic effects. There is a known strategy to mitigate systematic effects for neural nets but no firmly established procedure for BDTs. We present a method to incorporate systematic uncertainties into a BDT, the systematics-aware BDT (*saBDT*). We evaluate our method on open data of the ATLAS Higgs machine learning challenge and compare our results to neural nets trained with an adversary.

1. Introduction

In the past many particle physics experiments were limited in their accuracy by the lack of statistics. With the increase in luminosity the data samples are often so large that systematic uncertainties dominate the total error. If a multivariate analysis is employed, the systematics are typically not explicitly taken into account in the training and the method is optimized to maximize the performance estimate based on statistical uncertainty only. In the case of artificial neural nets (ANN) there exists a method to take into account systematic uncertainties in the training. Adversarial ANNs (AdvANNs)[1] are used in order to mitigate the effect of systematic uncertainties [2]. While this works well for neural nets, no firmly established method exists for Boosted Decision Trees (BDTs).

This paper attempts to fill this gap and presents systematics-aware BDTs (*saBDT*) which were developed during a Master thesis [3].[1] The goal is to sacrifice some statistical power while reducing the dependence on systematic effects.

We implemented our systematics-aware algorithm into BDTs with AdaBoost in TMVA of ROOT 6.14 [6]. Our example application is based on the public ATLAS data of the Kaggle Higgs Challenge [5]. This challenge provided us with the data used to train and evaluate our new method.

2. Evaluation

The impact of systematic uncertainties on a typical analysis is evaluated by comparing three datasets with different systematic effects. One set represents our best knowledge of the

[1] After our presentation at the conference we were made aware of a parallel development, QBDT, [4] which is similar to our method.

systematic effect (the standard dataset *std*) while the two others present the ± 1 σ variation of the effect (from now on called *variational datasets*). A systematic uncertainty on a quantity A is calculated by taking the value derived from the standard dataset and comparing it to the variational datasets. We define the average and the quadratic difference as:

$$\Delta_{sys}(A) = \frac{1}{2}\left(|A_{std} - A_{-1\sigma}| + |A_{std} - A_{+1\sigma}| \right) \quad (1)$$

$$\Delta_{sys2}(A) = \left((A_{std} - A_{-1\sigma})^2 + (A_{std} - A_{+1\sigma})^2 \right) \quad (2)$$

To quantify the influence of systematics on statistical results we use the **Advanced Approximate Median Significance** which was developed in the context of the Kaggle Higgs challenge. It considers both statistical and systematic uncertainties. A full explanation of the metric can be found in [7].

$$AAMS = \sqrt{2\left((s+b+b_{reg})\, ln\frac{s+b+b_{reg}}{b_0} - s - b - b_{reg} + b_0\right) + \frac{(b+b_{reg}-b_0)^2}{\sigma_B^2}} \quad (3)$$

$$b_0 = \frac{1}{2}\left(b + b_{reg} - \sigma_B^2 + \sqrt{\left(b + b_{reg} - \sigma_B^2\right)^2 + 4\,(s+b+b_{reg})\,\sigma_B^2}\right) \quad (4)$$

s is the number of signal correctly classified as signal by the BDT and b the number of background events wrongly classified as signal. σ_B is the difference between b obtained from the standard dataset and b obtained from the variational datasets: $\sigma_B = \Delta_{sys}(B)$. In addition we introduce a regularisation parameter b_{reg} to avoid excessive variations of $AAMS$ for small background values. Since the number of signal and background varies with the chosen BDT cut value, the maximum AAMS is taken as figure of merit.

The actual Kaggle Higgs challenge only used the **Approximate Median Significance** (AMS) which considers statical errors only. If the systematic error σ_B is negligible the $AAMS$ is equal to AMS with the same b_{reg} parameter. Aside from calculating a metric for the combined performance considering both statistical and systematic uncertainties, the absolute reduction of the impact of systematic uncertainties is also of interest. To estimate this we use the maximal σ_B obtained from varying the cut on the BDT value.

2.1. Systematics-Aware BDT

It is conceivable that data features which distinguish signal and background differ between standard and variational datasets thus creating a systematic effect in the BDT output. While identical performance between standard and variational dataset is unlikely to be achieved, the goal to reduce the performance differences between variational and standard datasets can be utilized in the systematics-aware training.

In the following, two procedures to introduce a penalty on performance differences will be discussed. In every step of tree building the performance is cross-checked with all three datasets. If the performances deviate from each other the tree building step is penalized by adding a penalty term to the quantity that determines the optimal node splitting. A similar penalty term will be added to the single tree Boost weight.

2.1.1. Penalty on the optimal node splitting When building a new tree at first the optimal split for the root node has to be found. All possible variables and split values are checked, the best one is chosen based on the *Gain*. The *Gain* can be calculated with multiple metrics. In this paper we worked with the Gini index G, which quantifies the separation power of a split. During the tree building, a split compares the metric of the mother node before the split to the metric

of the two daughter nodes after the split [6]. The Gini index for one node is defined through the purity p of a node which is 1 for a pure signal node and 0 for a pure background node:

$$p = \frac{N_{Signal}}{N_{Total}} \tag{5}$$

$$G = p \cdot (1 - p) \tag{6}$$

N_{Signal} is the number of signal events and N_{Total} the total number of events in the node. G follows an upside-down parabola, it has its maximum for $p = 0.5$ and is zero when $p = 0$ and $p = 1$. A purity close to 0 or 1 is wanted, as this would indicate a clean sample with just background or signal. Therefore the goal is to minimize the Gini index. This leads to the *Gain* being defined as:

$$Gain = G_{Parent} - G_{Left} - G_{Right} \tag{7}$$

With G_{Left} and G_{Right} defined as the Gini index of the left and right daughter node respectively. The *Gain* is maximal if the daughters' Gini indices are minimal, which therefore fits our requirements.

The *Gain* needs to be penalized for differing behavior on the variational datasets. To stay consistent the measure for the differing behavior is again based on the Gini index. The same split as on the *Standard* dataset is performed on the variational datasets and the Gini index of the daughter nodes on the variational datasets is calculated. The *Gain* is then penalized according to the quadratic systematic differences on the purity p:

$$NewGain = Gain - \frac{\lambda_{Cut}}{8} \cdot \sqrt{\Delta_{sys2}(p_{Left}) + \Delta_{sys2}(p_{Right})} \tag{8}$$

The difference is measured as quadratic sum of the differences of the two daughter nodes for all different datasets. λ_{Cut} is a penalty parameter to control the strength of the penalty and can be varied between 0 and 1.

In fig. 1 the $AAMS$ and σ_B^{max} of the *saBDT* depending on λ_{Cut} can be seen. For small λ_{Cut}

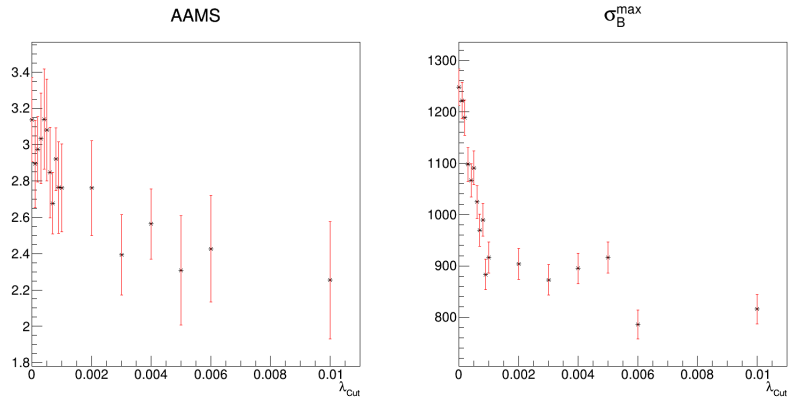

Figure 1. $AAMS$ (left) and σ_B^{max} (right) for different values of λ_{Cut}.

the $AAMS$ is stable, possibly with a slight increase. For $\lambda_{Cut} \gtrsim 0.001$ a drop in performance is visible. This happens along with a decrease of σ_B^{max}, which indicates the *saBDT* being more invariant against systematic shifts. The errors shown here are correlated. Therefore no clear conclusion can be drawn yet if the observed increase for low λ_{Cut} in $AAMS$ is meaningful.

2.1.2. Penalty on determination of the Boost weight The next step after building the decision trees of a BDT is the boosting. Every decision tree gets assigned a Boost weight BW which is used to calculate the final BDT scores. The weight of a tree is based on its performance and rewards good separation power. To assess the separation power the error rate err is used. This metric counts the number of misidentified events compared to the overall number of events and calculates the Boost weight:

$$err = \frac{N_{missid}}{N_{total}} \tag{9}$$

$$BW = \frac{1 - err}{1 + err} \tag{10}$$

This formula holds true for adaptive boosting, which is used in this study. The standard Boost weight is maximal for an error rate of zero and drops exponentially for higher error rates.

For the *saBDT* this weight should also take into account the different performances on variational datasets. In order to stay consistent the error rate is used as metric once more. The error rate of a single decision tree is calculated for all three datasets and compared. The boostweight is then multiplied with a factor depending on the difference in performance on the datasets.

$$NewBW = BW \cdot exp\left(-\frac{\lambda_{Boost}}{2} \cdot \Delta_{sys2}(err)\right) \tag{11}$$

From equation 11 is is clearly visible that the smaller the differences are, the closer the factor is to 1. Additionally there is a penalty parameter λ_{Boost} to tune the strength of the penalty. For 0 it gives the same behavior as a normal BDT, while for larger values the penalty for different performances increases.

3. Results

3.1. Systematic variation

As an example we consider a single systematic uncertainty on the jet energy scale that affects all jet quantities in the ATLAS machine learning challenge. We set the uncertainty to be ±1% relative uncertainty of jet transverse momentum and related quantities. Typical systematic uncertainties of that kind vary in ATLAS analyses from 1% to 4%.

3.2. Performance of saBDT

In figure 2 the $AAMS$ of the *saBDT* can be seen depending on λ_{Cut} and λ_{Boost} at the same time. For small values of both parameters the $AAMS$ increases, afterwards it falls off. This shows that the method works, even when both systematic controlling mechanisms are combined. Both parameters behave as expected; when increasing the penalty parameters, the systematic effects are reduced, although it also leads to a performance loss for larger penalty parameter values. The maximum AAMS is found to be at (λ_{Cut}, λ_{Boost})=(0.001,2). Performing an error analysis with the Bootstrap method [8] we find that the chance of this result being a statistical fluctuation is 17.9%.

3.3. Influence of strength of systematic effect

All prior analysis was done with a single systematic effect strength. Its value of 1% was chosen to be rather low. In table 1 the gain in $AAMS$ using the *saBDT* compared to a standard BDT is shown versus the systematic effect strength. The improvement due to the saBDT increases with increasing systematic strength. With small systematic effect the $AAMS$ is dominated by the statistical error and therefore a mitigation of the systematic error does not result in a large

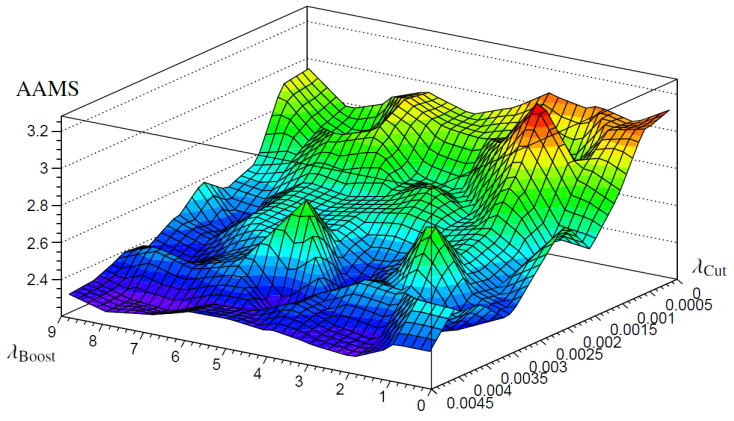

Figure 2. $AAMS$ for different values of λ_{Cut} and λ_{Boost}.

Systematic Variation	BDT ($AAMS$)	*saBDT* ($AAMS$)	% prob for stat. fluc.
20%	1.07±0.05	1.52±0.06	1.6%
10%	1.38±0.06	1.94±0.07	0.4%
3%	2.40±0.09	2.64±0.09	7.7%
1%	3.13±0.11	3.23±0.10	17.9%

Table 1. Achieved $AAMS$ for a standard BDT and the *saBDT* for different strength of the systematic effect as well as the probabilty for the improvement being due to a statistical fluctuation.

gain. In case the systematic effect is large, the $AAMS$ is dominated by the systematic error. The *saBDT* is able to mitigate this dominating factor and makes the loss in general performance neglectable. Therefore we conclude that the *saBDT* is better suited for use in areas dominated by larger systematic uncertainties.

3.4. Comparison between saBDT and AdvANN

In figure 3 the $AAMS$ results for the optimal saBDT and the optimal AdvANN are shown. Despite the curves peaking at different cut points both methods achieve similar results. The *saBDT* has a slight advantage and is able to achieve an $AAMS$ of 3.23 ± 0.10 in comparison to 3.08 ± 0.11 for the AdvANN (compared to 2.88 ± 0.09 for the regular ANN). No final statement can be made regarding which method works better. The *saBDT* may perform better in this study because the AdvANN is not as well-tuned, as indicated by the fact that the performance of the non-adversarial neural network is worse than the standard BDT.

As shown in [2] a well tuned ANN is expected to achieve a similar performance as a BDT. Therefore the worse performance of the AdvANN in $AAMS$ could be the result of the non-ideal tuning of the AdvANN. To get a clearer comparison between both methods more work has to be spent on optimizing both methods.

4. Conclusion

We present a new method to include systematic uncertainties in the training of BDTs, called saBDT. The saBDT performance is evaluated based on data of the ATLAS machine learning challenge equipped with a jet energy scale uncertainty. We find that for a large enough systematic variation the results of the saBDT are significantly better than those of a standard BDT. A

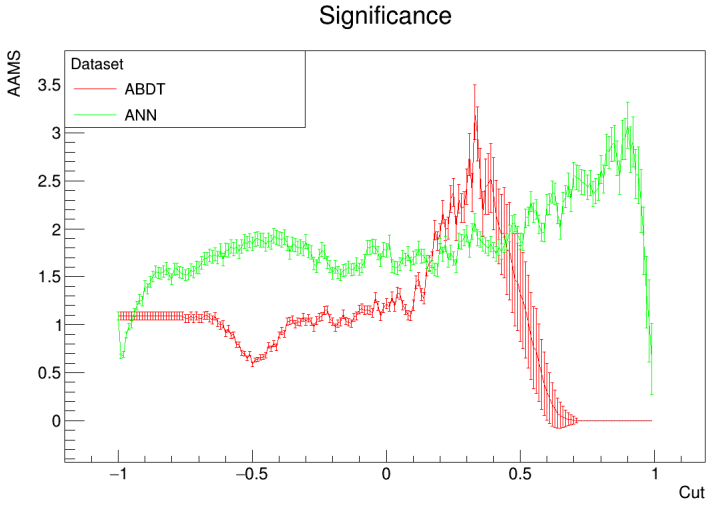

Figure 3. *AAMS* for *saBDT* and AdvNN vs cut value.

comparison to an adversarial neural net yields similar results but more studies are needed to determine which method works better.

Nevertheless this is overall a promising result as it shows the capabilities of BDTs. Being the most frequently used multivariate method at the LHC, it is a good sign that BDTs can still be improved and that they are capable of adapting to different tasks.

Acknowledgements

We would like to thank the conference organizers for a fruitful and splendid conference and Ruth Jacobs for a critical reading of the paper. During the conference we were made aware of another method to incorporate systematic uncertainties into a BDT [4].

References

[1] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y 2014 *Advances in Neural Information Processing Systems* pp 2672–2680

[2] Louppe G, Kagan M, Cranmer K 2017 *Advances in Neural Information Processing Systems* pp 981–990

[3] Alef T, Incorporation of systematic uncertainties in the training of multivariate methods, master thesis, BONN-IB-2019-04.

[4] Xia L G 2019 *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **930** 15–26

[5] Kaggle Higgs boson machine learning challenge documentation `https://higgsml.lal.in2p3.fr/documentation/` accessed 13-October-2018

[6] Hoecker A, Speckmayer P, Stelzer J, Therhaag J, von Toerne E, Voss H, *et al.* 2007 *arXiv preprint physics/0703039*

[7] Adam-Bourdarios C, Cowan G, Germain C, Guyon I, Kégl B and Rousseau D 2015 *NIPS 2014 Workshop on High-energy Physics and Machine Learning* pp 19–55

[8] Efron B, Tibshirani RJ: An introduction to the bootstrap, New York: Chapman & Hall, 1993