PROCEEDINGS OF SCIENCE

Pull-validation: A resampling method to improve the usage of low-statistics datasets

The IceCube Collaboration[†],

[†] *http://icecube.wisc.edu/collaboration/authors/icrc15_icecube*

E-mail: jan.luenemann@vub.ac.be

In high energy physics, many background dominated analyses suffer from limited statistics in simulation: With increasing efficiency of the event selection, the simulated samples are reduced so that in many cases the event number at final analysis level is very low. Due to limited computational resources, the production of more simulation is not always feasible. In these cases, it is helpful to extract more information from the available simulated datasets.
One way to deal with this issue in multivariate analyses (MVA) is by using resampling methods: The MVA is trained many times on small subsets that are randomly resampled from the complete dataset. The variation of the MVA output between the trainings can be interpreted as a probability density function (PDF) for each event. This PDF can be used to calculate a weight that is applied to each event instead of making a binary cut decision. With this procedure, events, that were normally removed by the event selection, can still contribute to the final dataset with a small weight. Another advantage is that pull-validation also provides an estimator for the uncertainty of the multivariate method. As an example of how the method can be used, we present a case-scenario from searches for physics beyond the Standard Model with IceCube.

Corresponding authors: J. Kunnen[1], J. Lünemann[*1], A. Obertacke Pollmann[2], F. Scheriau[3]

[1] *Vrije Universiteit Brussel*

[2] *Bergische Univertität Wuppertal*

[3] *Technische Univertität Dortmund*

The 34th International Cosmic Ray Conference,
30 July- 6 August, 2015
The Hague, The Netherlands

[*]Speaker.

PoS(ICRC2015)1211

1. Introduction

For typical analyses in high energy physics, very large datasets have to be evaluated. These datasets consist mostly of background, so that an effective event selection must be performed to remove parts of the data that presumably contain no signal. For this task, various multivariate analysis (MVA) tools can be applied, which assign to each event a value that can be interpreted as a probability that the event belongs to the signal or background class. An example are boosted decision trees (BDTs) [1], which assign a score between -1 (background-like) and $+1$ (signal-like) to each event. While the pull-validation technique described in this paper can be applied to different kinds of multivariate analysis methods, all examples provided in this papers are based on boosted decision trees.

In the first step of event selection, the classification algorithm has to be trained on labeled datasets (i.e. data with known class membership) to define efficient decision criteria exploiting differences in the measured variables of background and signal events. These datasets typically consist of simulated signal events and a combination of simulations of all known background types. After training, the decision criteria of the BDT are defined and the algorithm can be used to classify unlabeled data. An example of signal and background score distributions of a BDT is shown in Figure 1.

In analyses where the background rate is estimated from simulation, it can be problematic if the phase space of features for the background class is sparsely populated in the signal region. This will lead to low statistics for high classification scores so that the surviving background rate can be estimated only with large uncertainties. For cuts where no simulated background events survive, background estimation may be impossible. For example, a cut at 0.2 on the BDT score shown in

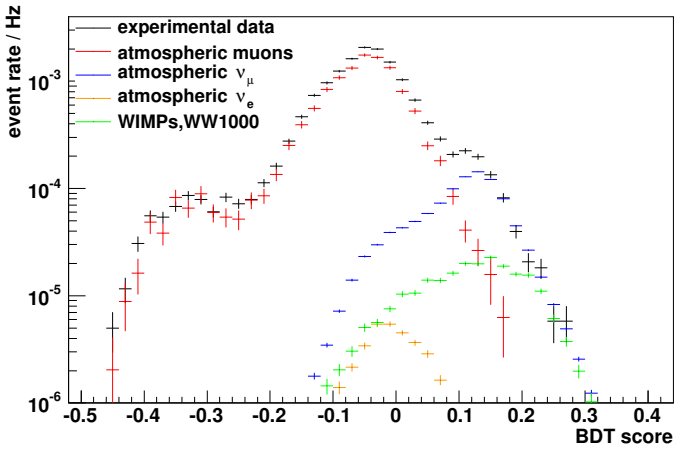

Figure 1: Score distribution for a single BDT, trained on a complete dataset available for training. This example is taken from an IceCube search for annihilating dark matter inside the Earth [2]. The goal of this selection is to achieve a sample with high neutrino purity. Experimental data, background simulations for atmospheric muons and neutrinos, and a neutrino signal from WIMPs with 1 TeV mass that annihilate inside the Earth into W^+W^- are shown.

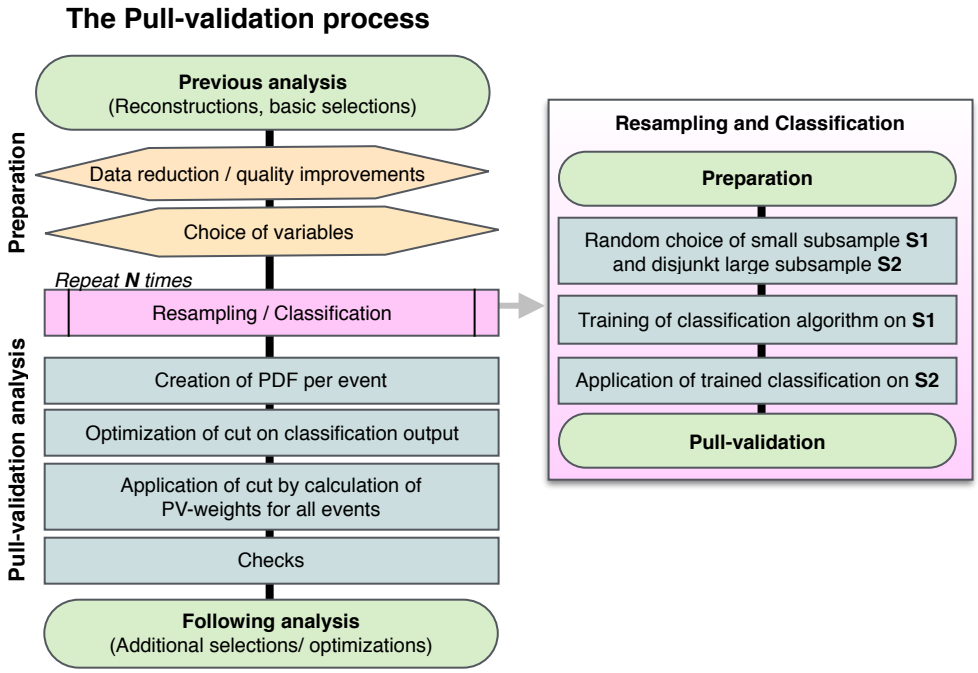

Figure 2: Schematic overview of the pull-validation technique in an analysis chain.

Figure 1 would remove all simulated atmospheric muons, falsely leading to a pure neutrino sample.

An obvious solution is the production of larger Monte Carlo datasets. However, simulation of large background statistics requires large computational efforts and is very inefficient if only the very signal-like tails of the score distributions are of interest. In contrast, the pull-validation technique utilizes the readily available datasets by enlarging the signal-like tail, as will be described in the following section. The term "pull" refers here to the selection of randomized samples and calculation of uncertainties. A schematic overview of this technique is given in Figure 2.

2. Pull-validation

A way to improve the usage of simulation statistics is to make use of uncertainties that can be derived by validation techniques. The most important validation techniques are cross-validation and bootstrapping [3, 4]. For cross-validation, the complete dataset is divided in N disjoint parts and the model is trained on subsamples composed of $(N-1)$ parts. After each training, the model is applied to the remaining subsample. In bootstrapping, events are randomly drawn and replaced for each subsample until these have the same size n as the original sample. Both methods achieve a smoothing of a distribution with insufficient statistics over the whole range. However, the effect is not large enough to compensate a lack of statistics larger than one order of magnitude. Pull-validation on the other hand can enlarge especially the tails of distributions.

Pull-validation includes a validation similar to bootstrapping. Like in bootstrapping, the variability of subsamples is used as an estimator for the variability of the whole sample. Instead of training the classification models with resampled sets of the size n, the subsamples are reduced

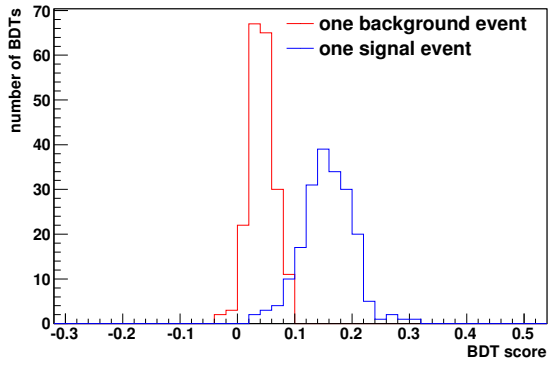

Figure 3: Score distribution of two example events, one of the background class and one of the signal class. As 200 BDTs were trained on different subsamples, the resulting scores for the same events show a variance.

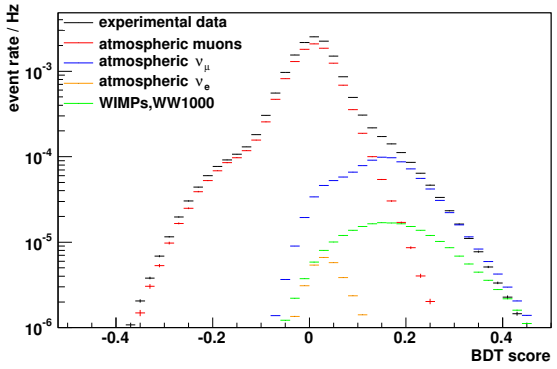

Figure 4: Sum of 200 BDT score distributions for the complete test dataset. Each BDT was trained on 10% of the training data. Compared to Figure 1, the distributions are smoother and broader.

in size (e.g. 10%), which leads to larger uncertainties. Since the subsamples are much smaller, the resampling can be done without replacement. This resampling and training is repeated several times (e.g. 200 times), each with a different subsample. This results in a large number of BDTs that will assign slightly different scores to the same event. The BDT score distribution of one event can be interpreted as probability density function (PDF), which is shown in Figure 3. The sum of all these PDFs (or equivalently the sum of the BDT histograms), shown in Figure 4, is considerably smoother that the single BDT distribution, shown in Figure 1. Note that the tails reach far deeper into the signal-like region.

A relation between pull-validation and kernel density estimation (KDE) is given by the fact that both methods smooth the score distributions. An important difference is that for pull-validation, the weights are derived from the uncertainties of the model while in KDE the smoothing is achieved without taking individual uncertainties into account. Calculating a weight for each event has the advantage that this information can be used in later stages after a cut on the BDT score.

In a classical event selection, a direct cut on the BDT score is applied, which means that an event is either accepted or rejected depending on whether its score is above a threshold or below. With pull-validation, this procedure can be replaced by calculating a weight for each event from its score distribution. This weight is determined by the fraction of scores above a threshold. This means that events with a medium score below the threshold can contribute to the final sample with a reduced weight, instead of being rejected completely. Therefore this procedure gives a higher statistics in the final sample. Figure 5 shows a comparison with a classical event selection, where a cut on a single BDT is applied, with an event selection using pull-validation weights. In this figure, the BDT cut was varied, which results in different final event rates. For each estimated event rate, the statistical uncertainty on the rate is clearly smaller when using pull-validation. Consequently, the distribution of any variable in the final sample will show a smoother behavior than without pull-validation.

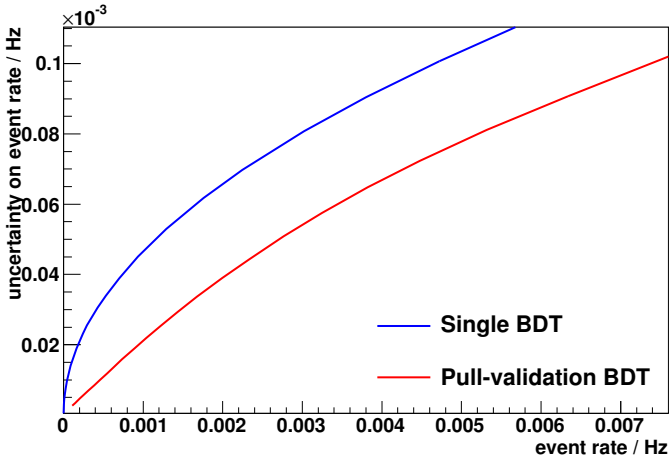

Figure 5: The uncertainty of the background rate versus the rate itself for varying cut values. The uncertainties were calculated as the square root of the sum of the squared weights. Using pull-validation, the uncertainties are significantly smaller than for a cut on a single BDT. The reason is that more events contribute to the final sample.

3. Manual

This section provides some information, that will help implement pull-validation in an analysis.

3.1 Prerequisites

Pull-validation can be used in combination with an MVA that is applied at any stage of an analysis, e.g. for preselections or the final event selection. Of course the computational effort is smaller at later stages, after some preselections have been applied to the data sets. However, some requirements must be fulfilled in all use cases. Then, ideally, pull-validation has the power to handle data with more than an order of magnitude missing simulation statistics. This was successfully tested in the applications described in the next chapter.

At first, the preselection of the events used for pull-validation and the BDT have to be tuned carefully in parallel. A critical issue is that pull-validation, as well as other methods, can not handle unknown (i.e. not simulated) event types. This can be checked by comparing the distributions of physical variables after the pull-validation using the calculated weights.

The interpretation of an MVA output is not straightforward. However, since the input variables have a physical meaning it can be interpreted with some experience. For example, a change of slope in a BDT distribution can be caused by an additional event type. This knowledge can be used to manipulate the result of pull-validation to strengthen its effect even further. This is for example useful if the available statistics for one type of background is much lower than for other background contributions. The event selection and the choice of variables for the BDT can be adjusted to strongly reduce the rate of this background type so that the other event types will compose the

dominating background. Consequently, it is not necessary to calculate the rate of the insufficiently simulated event type directly as this procedure assures that its contribution is negligible.

3.2 Choice of input variables

The variables for pull-validation have to be chosen carefully. The agreement between simulation and data must be excellent since small deviations might be multiplied due to the pull-validation procedure. The optimal number of variables may be higher than without using pull-validation. These training variables should have small correlations and describe the physics well, so that the output is reliable.

3.3 Definition of subsamples

For pull-validation, a comparably small subsample should be chosen (e.g. 10%). The small training sample leads to a large extrapolation of the tails of the BDT score distribution. In addition, this reduces correlation effects because the overlap of the subsamples is smaller. The number of subsamples has to be sufficiently large, so that additional repetitions would not change the result. For the analyses described in Section 4, the choice of 200 repetitions fulfills these requirements.

Once the pull-validation procedure has been applied, typically a cut on the result will be chosen. For the optimal choice of a cut value, a few criteria have to be kept in mind. The looser the cut is, the more reliable is the estimation achieved by pull-validation, i.e. the relative rate uncertainty is smaller. However, uncertainties of the order of 100% are manageable with this method. The number of events contributing at least partially to the final event rate decreases the statistical uncertainty. Another useful statistical check is how much every remaining event contributes to the rate if it had never been simulated. Pull-validation could be sensitive to systematic uncertainties due to the chosen variables which has to be checked and avoided.

3.4 Unblinding test

To prevent bias, the analyses of the IceCube collaboration are built on simulation. Only 10% of recorded data (the so-called burn sample) are typically used to validate the simulation sets. After the event selection is finalized, the remaining 90% of data are unblinded and used to calculate the result. As the optimization of the pull-validation is tested on the burn sample only, it is useful to check with a mock-unblinding if it is valid for the complete dataset. The burn sample is reduced to 10% of the original burn sample. The event selection, pull-validation and sensitivity calculation (including uncertainties) are finalized on this choice and then the other 90% of the sample are unblinded. This should give a rate in the calculated confidence interval, mostly near the estimated rate.

3.5 Overtraining check

Like for all supervised learning techniques, overtraining checks have to be performed when using pull-validation. In general, the risk of overtraining increases when reducing the size of the training data set. However, while individual BDT may be trained on statistical fluctuations, these effects will be averaged out by combining all BDTs. Figure 6 shows the score distributions of 200 BDTs for the complete training set and for an independent dataset. It can be seen that overtraining effects do not appear, if all BDTs are taken into account.

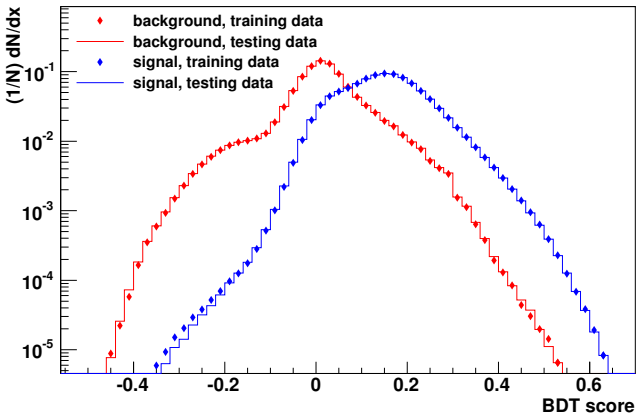

Figure 6: Overtraining test: the combined distributions of all 200 BDTs for the training set and for the testing set do not differ significantly. Possible overtraining in individual BDTs are averaged out.

3.6 Discrimination power

It is expected that the pull-validation technique does not suffer from a reduction in discrimination power compared to a classical training of a single BDT, if for both methods the same features and the same amount of data were used for training. This is indeed the case, as can be seen in Figure 7, which shows the background reduction versus the signal efficiency for a classical cut and for pull-validation event selection.

4. Applications to IceCube analyses

Pull-validation was already used in several analyses within the IceCube Collaboration. The method was first described and used in an IceCube unfolding analysis of the energy spectrum of atmospheric muon neutrinos [6]. In this analysis, a sample of very high signal purity is prepared by the event selection. The the quality of the unfolding that is based on this sample can be improved by smoother input distributions, which is achieved by introducing pull-validation to the analysis. The power of this method could be proven and a comparison with cross-validation showed the sanity of the results.

A second analysis is the search for magnetic monopoles [7], which would be a very rare signal in the IceCube detector. This analysis crucially depends on pull-validation and therefore the method was thoroughly checked as described in section 3. The final background rate after unblinding could be estimated successfully.

Finally, the plots throughout this paper are from a search for neutrinos from annihilating dark matter inside the Earth [2]. As this analysis relies on background simulations, a detailed understanding of systematic uncertainties is required. Therefore large enough statistics are necessary in the simulated data sets. The analysis benefits from the increased usage of background statistics by utilizing pull-validation, as shown in Figure 5.

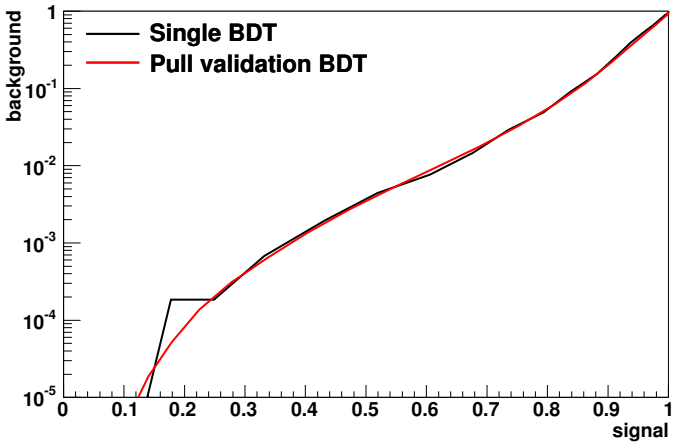

Figure 7: Receiver operating characteristic (ROC) plot. For a given signal efficiency, the background reduction is identical for the classical one BDT and for 200 BDTs. The step at hard cuts for the single BDT is due to low statistics for background estimation. This problem is reduced for pull validation.

5. Summary

Pull-validation is an efficient way to reduce the problem of sparse statistics. By estimating uncertainties of the scores and interpreting their distributions as a PDF, a weight can be calculated that replaces classical binary decisions of event selection. Consequently, more events contribute to the final sample and the statistical uncertainties are considerably reduced without reducing the discriminating power. Additionally, the estimated uncertainties can be taken into account as systematical errors.

References

[1] J. Quinlan, Mach. Learn., Vol.1, No.1 (1986), 81-106.

[2] IceCube Coll., PoS(ICRC2015)1205, these proceedings.

[3] P. Lachenbruch, M. Mickey, Technometrics, Vol.10,No.1 (1968), 1-11.

[4] B. Efron, Annals of Statistics, Vol 7, No.1 (1979), 1-26.

[5] A. Achterberg et al., Astropart.Phys. 26 (2006), 155.

[6] F. Scheriau, Dissertation, Universität Dortmund (2014).

[7] IceCube Coll., PoS(ICRC2015)1061, these proceedings.