Measurement/Simulation Mismatches and Multivariate Data Discretization in the Machine Learning Era

Mathis Börner[1], Tobias Hoinka[1], Maximilian Meier[1], Thorben Menne[1], Wolfgang Rhode[1], and Katharina Morik[1]

[1] *TU Dortmund, Dortmund, Germany;*
mathis.boerner@tu-dortmund.de

Abstract. In the era of machine learning and usage of high dimensional data many analyses in astroparticle physics face similar problems. This contribution presents approaches to face two typical challenges in the field: verify agreement between simulated and measured data and binning in multiple dimensions while preserving sufficient statistics. The presented approaches are based on widely used machine learning algorithms making them easy to use and providing good scalability in terms of size and dimensionality of the used data.

1. Introduction

With increasing complexity of measurements conducted in astroparticle physics analyses it is necessary to utilize as much information as possible. For this task machine learning techniques provide a powerful toolkit, and over the last years more and more of those techniques were used in the field. One of the most popular group of techniques are decision tree (Breiman et al. 1984) based classification and regression algorithms. These algorithms are supervised learning algorithms which means the model is trained with data for which the desired label is known.

The training of a decision tree is recursive procedure. In each step the dataset is split according to a cut in one of the observables. This cut is optimized to separate the data as good as possible. For the two split datasets the algorithm is called again. The recursion stops when a dataset can not be further split e.g. when an external condition is fulfilled or the events in the remaining dataset have the same label. Limiting the maximal depth of the tree or ensuring a minimal dataset size are often used break conditions. The stopping points are so called leafs and represent a sequence of cuts in the observables. A typical use case is the separation of signal and background events or the regression of a physical quantity like a particle energy. But the characteristics of decision tree based algorithms can be utilized even further.

2. Measurement/Simulation Mismatches

One of the key features of machine learning algorithms is that they can utilize many observables and especially their correlations among each other. In particle physics the labeled training data is most often generated from extensive simulations. To ensure the trained machine learning model can be applied on actual measured data the simulations have to be extremely accurate and in good agreement with the measured data. Validating the compatibility between simulated and measured data for many dimensions is especially challenging when the correlations between the features needs to be modeled correct. For this indispensable step no general approach is established.

The idea presented here is to train a classifier to separate between measured and simulated data (Martschei et al. 2012). For a perfect simulation the separation should be impossible and the classifier should not be better than guessing. If the classification is better than guessing, the trained model can be analyzed to identify observables with a significant mismatch in the simulation.

The approach is demonstrated with data from the IceCube experiment (Achterberg et al. 2006). After a first preselection each event is described by 333 observables. Those observables are used to train a Random Forest (Breiman 2001) to classify measured events (label = 0) and simulated events (label = 1). The resulting distribution for the different components are shown in Figure 1. The distribution for all features shows clearly that the classifier is not only guessing. The corresponding AUC is 0.685 ± 0.011. To identify the observables allowing for the separation the *feature importance* can be used. The *feature importances* state how often and how strong the observables were contributing to the training process. The sum of all *feature importances* is 1. For a purely guessing RF the distribution of the *feature importance* should be normal distribution located at $\frac{1}{N_{\text{Obs}}}$, because the model will be fitted to statistical fluctuations and all observables should contribute equally. In many applications some observables are used less often, because statistical fluctuations aren't as prominent in their distributions e.g. because they only have few discrete values. Therefore, the normal distribution can be slightly shifted. The definition we propose to identify outliers uses the median and the median absolute deviation (MAD). They are robust measures of location and variability of univariate samples. For normally distributed data the relation between the standard deviation σ and the MAD is $\sigma = 1.4826 \cdot \text{MAD}$. For the given IceCube example the classifications were carried out in a 10-fold cross validation and if in 8 out of 10 folds an observable had a *feature importance* greater than $\text{median} + 3 \cdot 1.4826 \cdot \text{MAD}$ they are considered as outliers. After the removal of the outliers the AUC for the classification is reduce to 0.537 ± 0.009 (see Figure 1). Remaining mismatches most likely come from uncertainties on true atmospheric muon flux.

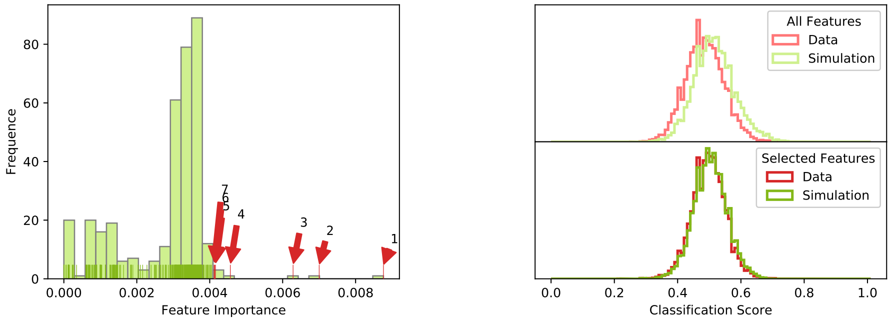

Figure 1. *Left*: Distribution of the *feature importances* for the classification with the full set of 333 observables including the seven outliers. *Right*: Distribution of the classification scores for the classification between measured and simulated events. For the full set of 333 observables (top) and for the final observables set (bottom).

3. Multivariate Data Discretization

To obtain physical parameters a binned likelihood fit is a very often used technique. In those fits a model is postulated which gives an expected number of events in the observable space. To fit the model parameters a likelihood is used in which the number of expected and measured events in the binned observable space is compared. The discretization of the observable space is a very crucial task, because it has to be fine enough to be sensible to the effects in data. But by using a fine binning the space can become very sparse. Low statistics in the bins can cause problems in the fit and lower the sensitivity. To find a binning optimized for the fit also the characteristics of decision trees can be utilized.

Our binning approach is demonstrated for an unfolding (Milke et al. 2013) measurement of the energy spectrum of a γ-source measured by the imaging air cherenkov telescope FACT (Anderhub et al. 2013). Goal of the unfolding is to obtain the energy distribution $\vec{f}$ of the measured γ-particles. The number of expected event in the observable space ($\hat{\vec{g}}$) is obtained via a linear model $\hat{\vec{g}} = \mathbf{A}\vec{f}$. The matrix $\mathbf{A}$ is the so called detector response matrix and consists of the conditional probabilities to measure an event in observable Bin(g_i) given that the particle is in energy Bin(f_j). The matrix is determined on simulations in which the sought-after energy of the γ-particle is known. To determine the energy distribution $\vec{f}$ the following likelihood is used:

$$\mathcal{L}\left(\vec{g}\middle|\vec{f}\right) = \prod_i \frac{\hat{g}_i^{g_i} e^{-\hat{g}_i}}{g_i!} = \prod_i \frac{\left(\mathbf{A}\vec{f}\right)_i^{g_i} e^{-\left(\mathbf{A}\vec{f}\right)_i}}{g_i!}.$$

With an ideal detector there would be a causal connection between observable and energy bin. Consequently, the detector response matrix would consist entirely of zeros and ones. For a realistic measurement the connection is far more ambiguous and tunning the observable binning is crucial to get keep this ambiguity as small as possible. The higher the ambiguity of the measurement the higher is the condition number κ of matrix $\mathbf{A}$ and the more ill-posed is the problem.

The idea of the decision tree based binning approach is to train a tree to classify the events in the different energy bins and to use the leafs of the tree for the binning. Every leaf of a tree represents a disjunct, rectangular area of the observable space and the whole space is covered by the leafs. As stated before, the binning needs to be a trade-off between a fine binning and preserving good statistic in each bin. Due to the training process the binning is explicitly optimized to separate between the different energy bins, which should help to make the problem as ill-posed as possible. To preserve reasonable statistic in each bin we can add external condition to the data to stop the training when less than k events are in a split dataset.

We use the condition number κ to compare equidistant binnings to binning using a decision tree. For all binnings it is ensured that each bin has sufficient statistics ($\geq$ 10 events). To validate that the condition number κ is a reasonable measure to compare different binnings an observable with low correlation to the energy is binned equidistantly. The resulting condition number is $\kappa = 90223.1$. Using a observable with a high correlation to the energy increases the condition number to $\kappa = 169.6$. Increasing the number of used observables for an equidistant binning is difficult when a certain number of events needs to be in every bin. For this example two observables with at least 10 events in every bin were feasible. Two highly correlated observable increase condition number to $\kappa = 84.6$. This example validates the expectation, that adding more informations lead to an less ill-posed problem. Using the same two highly correlated observables with the decision tree based binning only leads to a minimally increased $\kappa = 67.0$. The resulting binning is visualized in Figure 2. A way more significant improvement of the condition is achieved when using more than two observables. For 18 observables the condition number is $\kappa = 23.0$.

In Figure 2 an overview of the singular values of the different binnings is given. As demonstrated with the unfolding example it is possible to utilize a decision tree to obtain an optimized observable binning. The potential gain of this approach is highly dependent on the problem it is applied to. Its usability is not constrained to problems using a linear model. E.g. often in γ-astronomy a power-law $\Phi_0 E^{-\gamma}$ is used to model the energy spectrum. For such a model one could train the decision tree as a regression tree for the true energy and use the resulting model for the binning. Further potential extensions is to used boosting (e.g. AdaBoost (Zhu et al. 2009)) in the training and to use the boosted tree with the best condition.

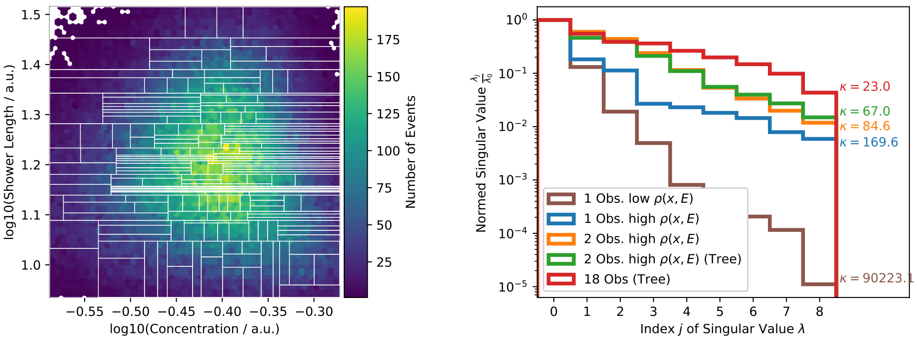

Figure 2. *Left*: Visualization of the decision tree based binning under usage of two Observables. *Right*: Singular values λ and condition κ for the FACT unfolding using different types of binning.

Acknowledgments. Part of the work on this contribution has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876. Special thanks to the IceCube and FACT collaborations for providing simulations and data.

References

Achterberg, A., et al. 2006, Astroparticle Physics, 26, 155

Anderhub, H., et al. 2013, JINST, 8, P06008

Breiman, L. 2001, Machine Learning, 45, 5

Breiman, L., et al. 1984, Classification and Regression Trees (Monterey, CA: Wadsworth and Brooks)

Martschei, D., et al. 2012, in Journal of Physics: Conference Series (IOP Publishing), vol. 368, 012028

Milke, N., et al. 2013, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 697, 133

Zhu, J., et al. 2009, Statistics and its Interface, 2, 349