EPJ Web of Conferences **251**, 03069 (2021) https://doi.org/10.1051/epjconf/202125103069
CHEP 2021

Basket Classifier: Fast and Optimal Restructuring of the Classifier for Differing Train and Target Samples

Anton Philippov[1,*] and *Fedor* Ratnikov[2,**]

[1]HSE University, Moscow, Russia

Abstract. The common approach for constructing a classifier for particle selection assumes reasonable consistency between train data samples and the target data sample used for the particular analysis. However, train and target data may have very different properties, like energy spectra for signal and background contributions. We propose a new method based on an ensemble of pre-trained classifiers, each trained of an exclusive subset, a data basket, of the total dataset. Appropriate separate adjustment of separation thresholds for every basket classifier allows to dynamically adjust the combined classifier and make optimal prediction for data with differing properties without re-training of the classifier. The approach is illustrated with a toy example. A quality dependency on the number of used data baskets is also presented.

1 Introduction

The common approach to event selection to boost the signal of background ratio in particle physics analysis assumes calibration and/or MC samples to train corresponding classifiers. Such a pre-trained classifier may be used in physical analysis later [1, 2]. This approach allows advanced training and validation of common purpose classifiers beforehand, and applying them to different physics analyses in a smooth and uniform way. However, this approach has some drawbacks.

Dependence on the training sample is one of the obvious problems of constructing classifier in the case of continuous spectrum. If in some parts of the spectrum there is a clear dominance of one type of analyzed objects, then the classifier would tend to assign this type to new objects only because there is such a big contribution in the training sample.

Another important example that prompts us to search for a new approach is the change in the spectrum in the target sample. The change entails a local change in the local class ratio, which we would like to reflect in the existing classifier. Consider the case when the spectrum of several features in the test sample has changed greatly compared to the target one. The problem is that we often can't just ignore these features (for example, exclude it with normalization). There are several reasons for this. First, by disregarding these features, we could lose some important information. Second, the indirect dependence on the other features can be significant enough, and its formal exclusion from the set of features will not lead us to the desired result. This problem can be solved by dividing the spectrum into regions (called baskets) so that each part has its own classifier. Let's divide the range of values of

*e-mail: anton.philippoff@gmail.com
**e-mail: fedor.ratnikov@cern.ch

each of the features into equal intervals. Thus, for N features, we will divide the area of their values into a certain number of N-dimensional baskets.

We consider the case of working with a metric ROC AUC. If we deal with another sample of analyzed objects, which forms a different spectrum shape than we have in the training sample, we can thus choose thresholds in each basket in order to maximize the signal level for a given background level.

It seems that the most natural approach to the classification problem in the case of continuous spectra would be to recreate the spectrum of the observed data on the Monte Carlo data, and train the classifier on the latter. But this is a rather complicated procedure. In this article, we will consider an alternative approach, where the classifier is trained only once, adapting to new data using an optimization procedure.

The motivation for this work is to build a classifier that has two properties:

- tolerance to changes in the spectrum.
- the ability to quickly recalculate the parameters of the classifier when the spectrum changes (without re-training the classifier). The metric that we are interested in is efficiency for a given background level.

As we suggested earlier, let's divide the spectrum into a number of baskets. For each basket we build its own classifier, that maximizes the area under the ROC curve for data in this basket. Due to narrowness of the basket, the classifier trained for each one only marginally depends on the particular training full spectra shapes. Thus, due to the general tolerance of ROC AUC to imbalance of classes, the obtained classifiers trained for each basket become tolerant to possible significant variations of distributions for the full training data samples.

Now we can solve the problem of maximizing the signal level for a given background level; in other words, to do this, we need to select a cut-off threshold in each basket so that for a given amount of background events across all baskets, the sum of signal events is maximum, i.e. solve the optimization problem.

We can immediately draw attention to the drawback of the method: to solve the optimization problem, we need to know the ratio of signal and background events between the baskets. The main advantage is the ability to construct a set of classifiers once and then manage only by setting thresholds in each of them, solving the optimization problem.

2 Optimization problem

This study is inspired by necessity to train the signal-background classifier on available calibration data samples and further apply it to physics analysis with energy distributions for both signal and background, that are very different from the training samples. The classifier construction procedure for such a case is like the following: the energy range is divided into several (7 in our case) baskets of equal size, for each of which we construct separate classifier. We use XGBoost to train these classifiers. XGBoost hyper-parameters are selected separately for each basket. As a result, for each basket we obtain a fine-tuned classifier accompanied by the ROC curve. Obtained ROC curves are approximated then by a 12th degree polynomial.

The full classifier is then build as a set of classifiers for each basket by selecting a set of separately optimized working point thresholds for each basket, that all together provide a necessary signal-background separation for the full classifier. Note, that individual thresholds, that are working points on every ROC curve, are now driven by the energy distribution of the actual sample to which this classifier is applied to. Thus the optimization problem is to find such a set of working points for every basket which would optimize the separation power of the full classifier on the dataset with given energy distribution.

In this case we deal with the optimization of a convex function on a convex set, so the optimization procedure is quite simple. Let's look at the optimization problem:

$$\begin{aligned} \min_{\xi} \quad & \sum_{i=1}^{N} m_i \xi_i \\ \text{s.t.} \quad & \frac{\sum_{i=1}^{N} f_i(\xi_i) n_i}{\sum_{i=1}^{N} n_i} = \alpha \end{aligned}$$

Where:

m_i - number of background events for the i-th basket,

n_i - number of signal events for the i-th basket,

ξ_i - percentage of signal events for the i-th basket,

α - level of a signal.

N - number of baskets.
To solve this problem, we need to solve the regression problem for each ROC-curve within each section.
Optimization procedure includes 3 steps:

- the projection of the gradient onto the tangent plane

- lowering the vector to the surface of constraints

- repeating the first 2 steps until convergence

3 Toy example

To test the approach, let's apply it to a toy example. We consider the following model: we parameterize 2 families of 3-D distributions (we can think of distribution 1 as a background and distribution 2 as a signal)

Distribution 1: two 2-D Gaussians with centers at $(-E, E)$ and $(E, -E)$ and fixed variance, with some distribution for E.

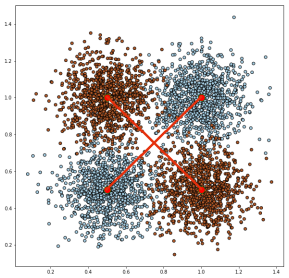
Distribution 2: two 2-D Gaussians with centers at (E, E) and $(-E, -E)$ and fixed variance, with some different distribution for E.

Our goal is to train the procedure on the data sample with some distributions for the parameter E (train sample A), and then apply the classifier to the data sample with another distribution of the parameter E (test sample B). The quality of the resulting classifier is evaluated by comparing the obtained result with the results of 2 other classifiers:

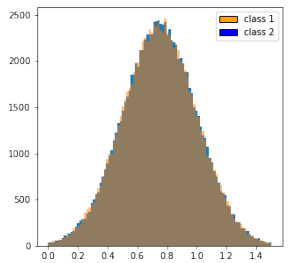
1 a plain classifier trained on the sample A, with performance evaluated on the sample B (worst case scenario).

2 a plain classifier trained on the sample B with performance also evaluated on test sample B (best case scenario).

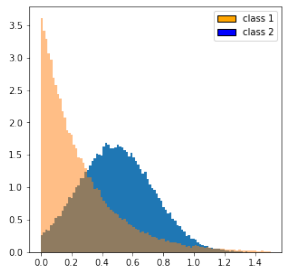
In this toy example, the sample A is constructed using identical Gaussian distribution for both signal and background. For the sample B, E has the same Gaussian distribution for the signal, but the background has an exponential distribution for E.



(a) Toy sample. E is a distance represented by red lines



(b) E distribution for the sample A



(c) E distribution for the sample B

Figure 1: Toy example distributions

3.1 Results

To demonstrate the relationship between the number of baskets and the quality of the basket-based dynamic classifier we consider ROC curves four scenarios:

- ROC-curve for the classifier trained on the sample A and being tested on the sample B (plain classifier baseline)
- ROC-curve for the classifier trained on the sample B, being tested on the sample with the same distribution of the B (ideal case)
- basket classifiers with 3 baskets trained on the sample A and being tested on the sample B
- basket classifiers with 7 baskets trained on the sample A and being tested on the sample B

We expect to see that ROC-curves for both basket classifiers are located between the baseline and ideal ROC-curves, and the performance of the 7-basket classifier would be better than the performance of the 3-basket classifier.

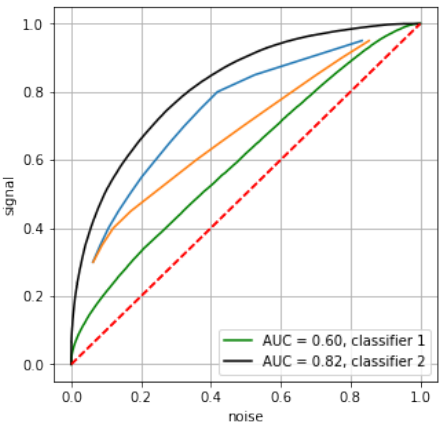


Figure 2: green line - efficiency for case (1), black line - efficiency for case (2), orange line - basket classifier with 3 baskets, blue - basket classifier with 7 baskets

Figure 2 demonstrates a dependency of the signal efficiency versus background rate for these cases. Greenline - efficiency for case (1), black line - efficiency for case (2), orange line - basket classifier with 3 baskets, blue line - basket classifier with 7 baskets. As it was expected, basket classifiers are located between cases (1) and (2), and their quality improves with the number of baskets into which we divide the spectrum.

4 Summary

In this paper, we present a concept of basket-based dynamic classifier. Such classifier demonstrates a tolerance to significant variations of the spectrum of the target analyzed data from data used for training. The procedure of fast adjustment of the basket-based classifier for a given analysis performance is also shown. In case of real experiments, such a basket-based classifier may be trained and validated only once in advance of data analyses. Further adjustments to real spectra of particular data analyses does not require re-training if using *a priori* knowledge of shapes of the target data sets.

References

[1] V. Chekalina, F. Ratnikov, *Machine Learning approach to $\gamma/\pi 0$ separation in the LHCb calorimeter*, Phys.: Conf. Ser. 1085 042036

[2] M. Calvo, E. Cogneras, O. Deschamps, M. Hoballah, *A tool for $\gamma/\pi 0$ separation at high energies*, LHCb-PUB2015-016.

[3] Tianqi Chen, Carlos Guestrin, *XGBoost: A Scalable Tree Boosting System*, arXiv:1603.02754 [cs.LG]