Machine Learning approach to boosting neutral particles identification in the LHCb calorimeter

A Boldyrev[1], V Chekalina[1], F Ratnikov[1,2] on behalf of the LHCb collaboration

[1] National Research University Higher School of Economics, Laboratory of Methods for Big Data Analysis, 3 Kochnovsky Proezd, Moscow 125319, Russia
[2] The Yandex School of Data Analysis, 11/2 Timura Frunze St., Moscow 119021, Russia

E-mail: alexey.boldyrev@cern.ch

Abstract. We present a new approach to identification of boosted neutral particles using Electromagnetic Calorimeter (ECAL) of the LHCb detector. The identification of photons and neutral pions is currently based on the geometric parameters which characterise the expected shape of energy deposition in the calorimeter. This allows to distinguish single photons in the electromagnetic calorimeter from overlapping photons produced from high momentum π^0 decays. The novel approach proposed here is based on applying machine learning techniques to primary calorimeter information, that are energies collected in individual cells around the energy cluster. This method allows to improve separation performance of photons and neutral pions and has no significant energy dependence.

1. Introduction

A few important analyses at LHCb need to reconstruct energetic photons in the final state. The notable example is the analyses of radiative B decays which serve as a sensitive probe for various extensions of the Standard Model. Such decays are affected by the backgrounds from B decays where a photon is substituted by a neutral pion (π^0). Since the π^0 can be produced with high momentum in the laboratory frame, the two photons from its decay can produce overlapping clusters in the calorimeter, and can thus be misinterpreted as the single energetic photon. It is thus very important to be able to distinguish between the high-momentum photons and π^0 using the shape of the calorimeter clusters. This paper describes the novel approach to identification of boosted neutral particles, evaluation of its performance obtained on MC samples, and discusses specific issues when transferring discriminative models from simulation to real world.

2. Electromagnetic Calorimeter

The LHCb calorimeter system performs several tasks, providing the first level trigger with high transverse momentum photon, electron and hadron candidates, measuring their energies and positions and performing the separation between photons, electrons and hadrons [1]. The LHCb ECAL [2] is based on the Shashlik technology of alternating scintillating tiles and lead plates, preceded by a Preshower (PS) and contains 1536/1792/2688 cells in its inner/middle/outer regions, respectively.

When the ECAL cell has an excess in energy deposition (compared to the adjacent cells) it will originate a cluster according to the following procedure [3]. Energy deposits in ECAL

cells are clusterised applying a 3×3 cell pattern around the local maximum of energy deposition or seed cell. Consequently, the seed cells of the reconstructed clusters are always separated by at least one cell. The transverse energy of the seed cell is required to be larger than 50 MeV. Neutral clusters are identified as those clusters that do not match to charged tracks extrapolated to the calorimeter surface. For each track-cluster pair, χ^2_{2D} is obtained taking into account: the position of the point of intersection of the extrapolated track from the calorimeter, the covariance matrix of track parameters, the position of the barycenter of the cluster and the matrix of the second moments of the cluster.

The signature of a π^0 decay in the ECAL depends on the kinematics of the two photons. Low-momentum π^0 produces two separated clusters. Such π^0 are classified as resolved π^0 and its reconstruction is based on the invariant mass of the photon pair. High-momentum π^0 produces a single cluster and it is classified as a merged π^0. The transition region between high- and low-momentum π^0 occurs around 2 GeV/c.

The discriminating features separating photon and merged π^0 clusters are based on the cluster shape. Merged π^0 clusters are expected to be elongated and asymmetric due to residual offset between two photons, while genuine photon clusters are expected to be more symmetrical. Full description of the variables used for the discrimination between photons and merged π^0, which we call shape-based approach (or *baseline*), can be found in [3].

3. Current Machine Learning approaches to Particle Identification

Particle identification algorithms can be based on multivariate classifiers which allow a straightforward combination of information originating from different subdetectors (i.e. PS and ECAL) into a discriminant output. Specifically to our problem, it allows extending the baseline method to consider track-cluster matching or the shape of the neutral cluster.

Particle identification algorithms are trained to separate photons signatures from hadrons and electrons, which may have passed the neutral cluster selection, and high-energy π^0. The classifier output of both baseline and XGBoost approaches is displayed on Figure 1 and 2 using simulated data samples for both training and performance evaluation. The performance of both approaches is presented as the dashed line in Figure 3.

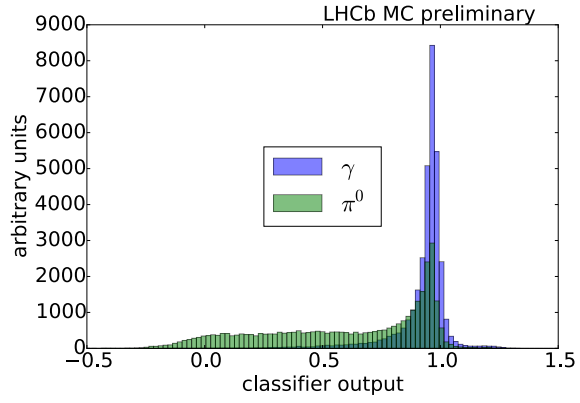

Figure 1. Baseline output (IsPhoton variable)

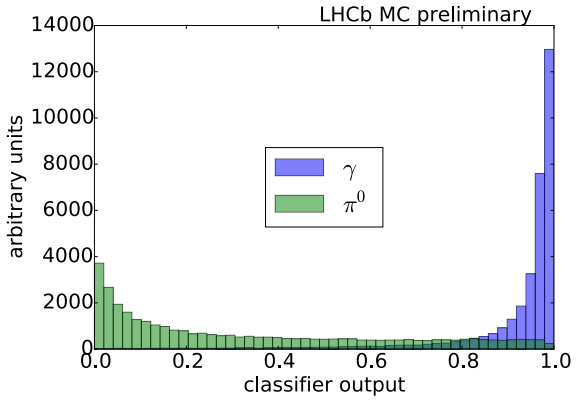

Figure 2. XGBoost approach output

Due to MC/data disagreement in the input variables used to build the γ/π^0 separation variable, discrepancies are also expected in the output of the method. To get a true estimate of the selection efficiency for a given cut on the classifier output, calibration samples from real data are needed.

4. Calibration samples

In order to calibrate the performance of the discriminative variable, $B^0 \to K\pi\gamma$ reconstructed events are used as calibration samples for photons and $B^0 \to K\pi\pi^0$ for π^0. The kinematically similar decay is chosen to prevent a biased output with respect to the particle energy. To prove stability of the method, additional π^0 samples from $B^0 \to J/\psi K^*$ (where $K^* \to K\pi^0$) are used. Due to the high muon trigger efficiency, the presence of J/ψ mesons decaying into pair of muons provides high detection efficiency of the selected decays of B^0 mesons. We required K and π to have transverse momentum $p_T > 500$ GeV/c and the vertex quality of the K and π tracks forming the K^* candidate to be $\chi^2(K^*) < 9$. We considered energy clusters with transverse energy $E_T > 2$ GeV.

5. New approach

To separate photons and merged π^0, we take into account energies in a 5×5 ECAL window and the PS cells around the cell seed. However we do not build any sophisticated features based on physics considerations, but rather consider plain values of energy allocated in every cell of the 5×5 matrix in both ECAL and PS as features. These are 50 plain features to be used by the classifier. The technical aspects of the proposed approach was discussed in detail in the conference paper [4]. In this study, the inner region of the ECAL is used as a reference for quantitative comparison of the classifiers.

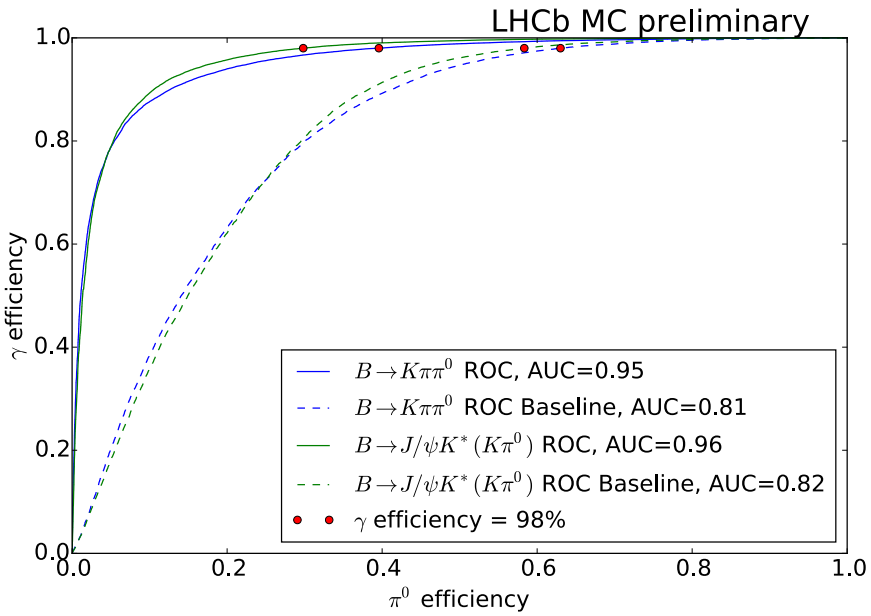

Figure 3. Receiver Operating Characteristic (ROC) curves for the baseline (dashed line) and new approach (solid line). Different colours refer to different test samples

6. Classifier selection and performance

We used several classifiers based on Neural Network and Boosted Decision Tree architectures.

For possible NN architecture, we build two parallel branches fed from preprocessed information from the ECAL and PS. We found that using PS branch with smaller number of units together with ECAL branch could improve quality of the classifier. Among the hyperparameters for the NN are: number of layers and number of units in each layer, method of regularisation,

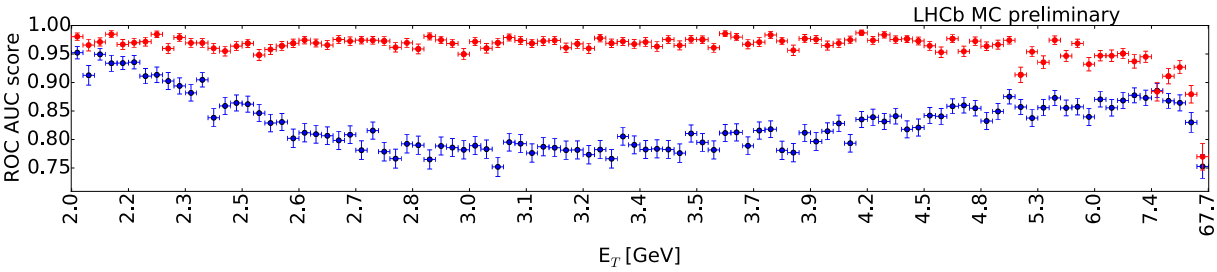

Figure 4. Baseline (blue) and BDT approach (red) model quality as a function of transverse energy

activation function. To achieve best quality of the classifier, we tuned all the hyperparameters leaving the network topology unchanged. To determine the number of training epochs, we monitored the train-test score curve and observed when it reached the plateau. As a result, we used 3-fold cross-validation and trained the network up to 5000 epochs. The insufficiently complicated structure of input features leads to a degradation of classifier quality with NN configurations with 3 and 4 hidden layers.

For the BDT approach, we assayed XGBoost, CatBoost and LightGBM classifiers. For each classifier, we find the best configuration by tuning the boosting parameters using ModelGym [5]. The different classifiers qualities are found to be very close to each other.

As BDT-based classifiers demonstrate better performance for the problem than the best NN configuration, we choose the BDT approach. The default XGBoost [6] was selected (estimators = 2000, learning rate = 0.05, max depth = 5, min child weight = 2). We used XGBoost with 6000 trees with depth set to 3. Figure 3 demonstrates the performance of the new approach in comparison to the baseline one. The score in the ROC curve improves from 0.89 for the baseline to up to 0.97 in the new approach. Considering 98% photon efficiency, the new approach reduces fake rate from about 60% to about 30%.

An unbiased behaviour with respect to energy is an important characteristics of the classifier, as it directly affects systematic uncertainties for physics analyses. As one can see in Figure 4, the new approach displays a flat efficiency profile with respect to E_T.

7. Conclusion

We developed a new procedure to separate photons from merged π^0. The new approach shows good performance on simulated data and reasonably good performance using only the ECAL energy deposits. The proposed classifier shows negligible energy dependency which can be useful to estimate systematic uncertainties for physics analyses.

However, careful consideration of the simulated data used in efficiency evaluations should be taken. Since the new approach uses solely energy deposits in the ECAL, it can be considered as a necessary part of future neutral particle identification tools.

8. Acknowledgements

The research leading to these results has received funding from Russian Science Foundation under grant agreement n° 19-71-30020.

References

[1] LHCb collaboration, A. A. Alves Jr. *et al.*, *The LHCb detector at the LHC*, JINST **3** (2008) S08005
[2] LHCb collaboration, R. Aaij *et al.*, *LHCb detector performance*, Int. J. Mod. Phys. **A30** (2015) 1530022, arXiv:1412.6352

[3] M. Calvo, E. Cogneras, O. Deschamps, M. Hoballah, *A tool for γ/π^0 separation at high energies*, LHCb-PUB-2015-016
[4] V. Chekalina and F. Ratnikov, *Machine Learning approach to γ/π^0 separation in the LHCb calorimeter*, Phys.: Conf. Ser. **1085** (2018) 042036
[5] Model Gym Project 2017 Model Gym [software] Available from `https://github.com/yandexdataschool/modelgym` [accessed 2019-05-14]
[6] XGBoost: A Scalable Tree Boosting System. `arXiv:1603.02754 [cs.LG]`, March 2016