2 MACHINE LEARNING APPLICATION FOR PARTICLE IDENTIFICATION IN MPD

INTRODUCCTION

KEY CONCEPTS

MACHINE LEARNING

Machine learning in the case of the MPD experiment, machine learning methods, such as gradient boosting on decision trees (CatBoost), are used to enhance particle identification in regions where conventional methods fail, processing features from experimental data to correctly classify particle types.

CATEGORICAL BOOSTING (CATBOOST)

CatBoost in the MPD experiment, is employed to address the particle identification problem, using subdetector data to correctly classify particle types across different momentum ranges, demonstrating greater efficiency than conventional methods under certain conditions.

PARTICLE IDENTIFICATION (PID)

Particle Identification is the process of determining the nature and characteristics of individual particles detected in a particle physics experiment, using information obtained from specific detectors such as the Time-Projection Chamber (TPC) and Time-of-Flight (TOF) detector.

PARTICLE IDENTIFICATION IN MPD EXPERIMENT

WHAT IS UNDERSTOOD?

Particle identification (PID) in the MPD experiment relies on the TPC and TOF subdetectors, whose details are provided in specific technical reports. These subdetectors enable achieving the required identification by reconstructing trajectories and measuring properties such as momentum, charge, energy loss, squared mass, number of TPC hits, pseudorapidity, distance of closest approach, and vertex coordinates for six particle species: proton, positive and negative kaons, positive and negative pions, and antiproton.

The PID problem was approached as a multiclass classification task in machine learning, where a CatBoost classifier trained on three Monte Carlo datasets (prod01 with real distribution, prod04 and prod05 with uniform distribution) was used. These datasets were generated simulating Bismuth and Bismuth Bi + Bi collisions at 9.2 GeV, expected to be the first collision systems in MPD. Testing data was generated under the same conditions as prod05, and CatBoost parameters were tuned using the Tree-Structured Parzen Estimator algorithm in Optuna, following developer recommendations.

This study demonstrates that the CatBoost algorithm is effective in improving accuracy in particle classification in the MPD, excelling in identification at critical momentum ranges where conventional methods show limitations.

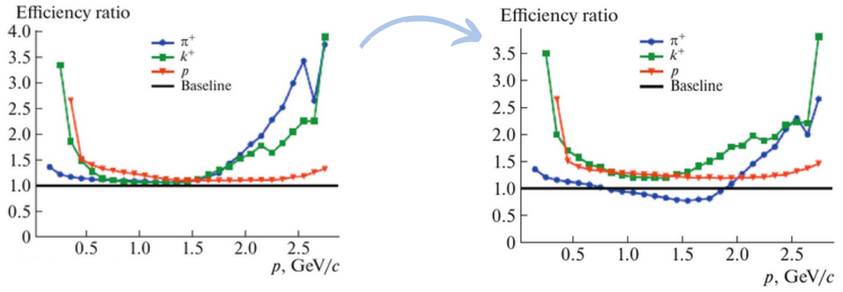
OUTCOMES

EXPLORAMOS LA SITUACIÓN ACTUAL DE LA COMPAÑIA

This section demonstrates the feasibility of using CatBoost for particle identification (PID) in the MPD experiment. Two metrics were used to evaluate model performance: E^s (efficiency) and C^s (contamination) for each particle species s. Efficiency E^s indicates how often the classifier provides correct answers, while contamination C^s describes the proportion of incorrectly classified tracks among those identified as type s.

$$E^s = \frac{N^s_{\text{corr}}}{N^s_{\text{true}}}, \qquad C^s = \frac{N^s_{\text{incorr}}}{N^s_{\text{corr}} + N^s_{\text{incorr}}},$$

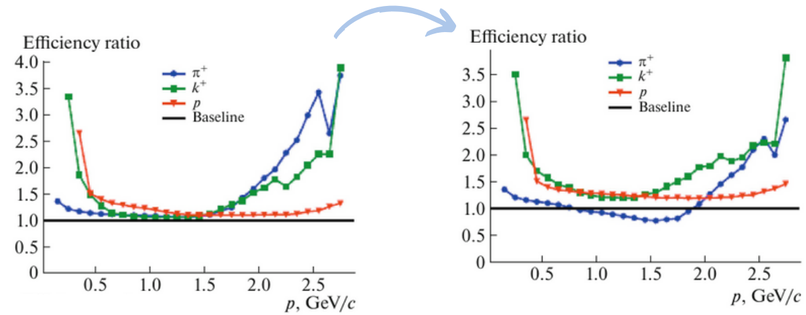
The PID performance is depicted in Figure 1, showing how particle identification efficiency varies across different productions from prod01 to prod05 for middle and high momentum values. Proton efficiencies increase by an average of 10%, whereas the efficiency and contamination of positively charged pions decrease by approximately 20%.



Notably, protons outnumber positive pions sevenfold when p>1.5p > 1.5p>1.5 GeV/c, affecting classification dynamics. Overall classifier efficiencies on test data are approximately 96.48%, 96.71%, and 95.99%, indicating robust performance, particularly in distinguishing between protons and pions across different production scenarios.

Similar trends are observed for negatively charged pions, where identification efficiency decreases as production shifts from prod01 to prod05, albeit with significantly lower contamination. Conversely, antiproton contamination exceeds that of protons, highlighting challenges in distinguishing these particle types.

CatBoost classifiers were compared with the n-sigma method currently used at MPD, as shown in Figure 2, providing a clearer assessment of the obtained results and confirming CatBoost's effectiveness in enhancing particle identification compared to conventional methods.



CONCLUSION

The conclusion of the study highlights that CatBoost demonstrates promising efficiency in particle identification in the MPD experiment, outperforming the currently used n-sigma method in several cases. Specifically, CatBoost showed better results in extreme momentum ranges ($p < 0.7$ GeV/c and $p > 1.5$ GeV/c), where the n-sigma method has significant limitations or lacks efficiency. This finding suggests that machine learning methods like CatBoost are suitable and offer substantial improvements in particle identification accuracy compared to traditional approaches used in particle physics. Furthermore, the study proposes continued exploration of additional particle characteristics to enhance model stability and performance across various experimental conditions of the MPD detector.