Domain Adaptation Techniques in Particle Identification for the ALICE experiment

Michał Kurzynka[1], Kamil Rafał Deja[1], Georgy Kornakov[2] and Tomasz Piotr Trzciński[1]

[1] Institute of Computer Science, [2] Faculty of Physics, Warsaw University of Technology

Abstract

- Quality of standard machine learning models is limited by quality of mapping between simulation and real data.
- Current methods of training the machine learning based classifiers are limited only to simulation data, where the ground truth, i.e. true particle pdg code is known.
- Models that are trained in this fashion and then applied to real data, which has different distribution, tend to give poor results.
- We propose to use domain adaptation model, which is able to find common features between simulation and production data sets.

Introduction

Classifying particle types on the basis of detectors response is a fundamental task in the ALICE experiment.

Methods currently employed in this job are:

- Based on linear classifiers,
- Built using only simulation data due to lack of labels (*pdg code*) in case of production data
- Requierd to be fine tuned to match production data set distribution.

We propose domain adaptation model based on artificial neural networks, which will be able to utilitize both simulation and production data during training process and correctly classify particles basing only on production data.

Data description and preparation

- We gathered over 300000 samples of both production and simulation ESD data.
- Using tree boosting model, we were able to automatically determine six most important features using SHAP values.

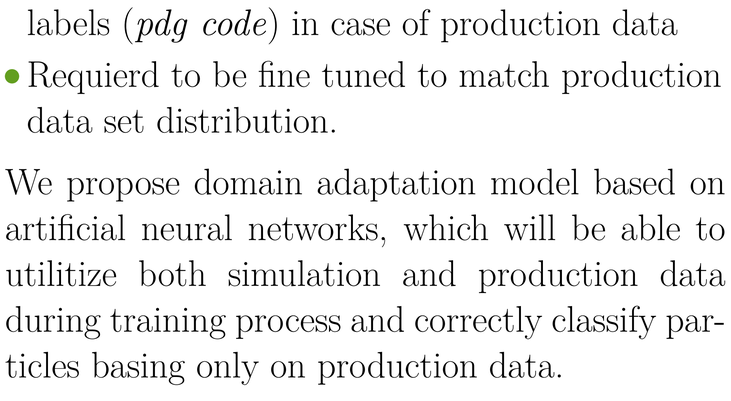

Figure 1:Domain shift between production and simulatin data, *TPC signal*

Model architecture

We propose one versus all based model based on Domain Adversarial Training of Neural Networks. Architecture consists of three neural networks:

- Feature mapping network, which maps features of both data sets into common, domain invariant latent space.
- Particle classification network, which classifies particles basing on domain invariant latent space.
- Domain discriminator network, which classifies domain of each particle.

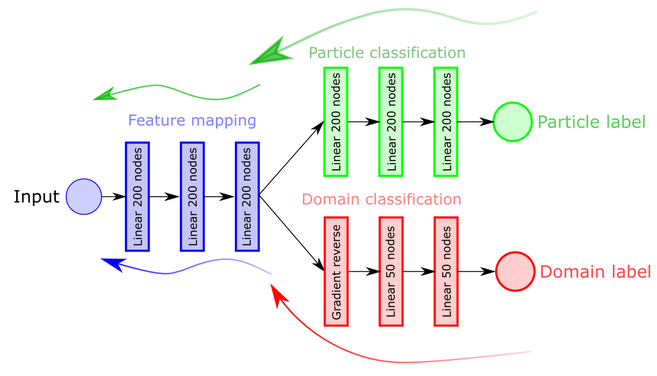

Figure 2:Architecture of proposed adaptation model.

Training process

- Pre-train both praticle classifier and domain discriminator on half of simulation data samples.
- Propagate mixed batches of remaining simulation data samples with production data samples.
- Split each batch, propagate only simulation data through particle classifier.
- Propagate both simulation and production data through domain discriminator.
- During backward propagation reverse domain loss function gradient to maximize domain discriminator confusion.

Results without domain adaptation

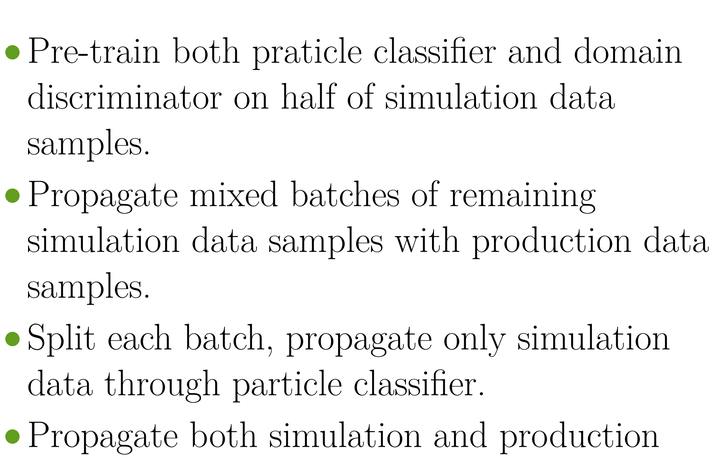

Figure 3:Particles classified by model without domain adaptation as electrons, *TPC Signal* histogram.

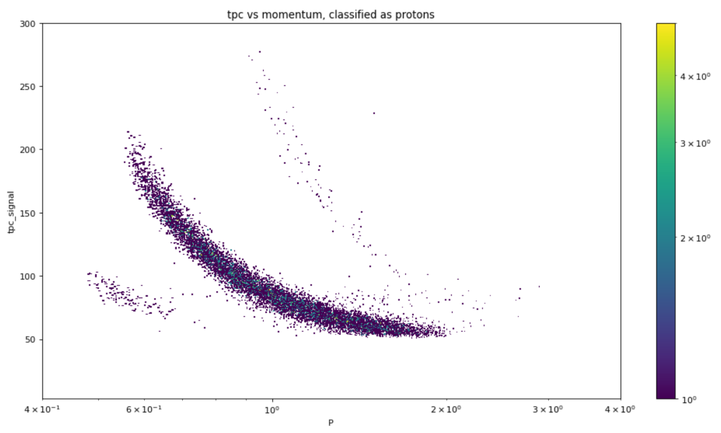

Figure 4:Particles classified by model without domain adaptation as protons, *TPC Signal* histogram.

Results with domain adaptation

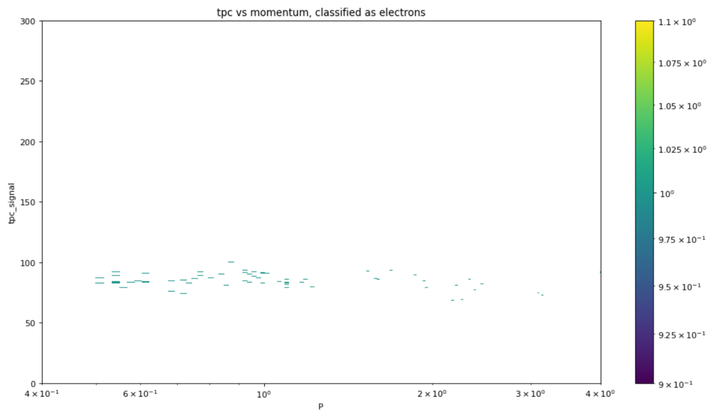

Figure 5:Particles classified by model with domain adaptation as electrons, *TPC Signal* histogram.

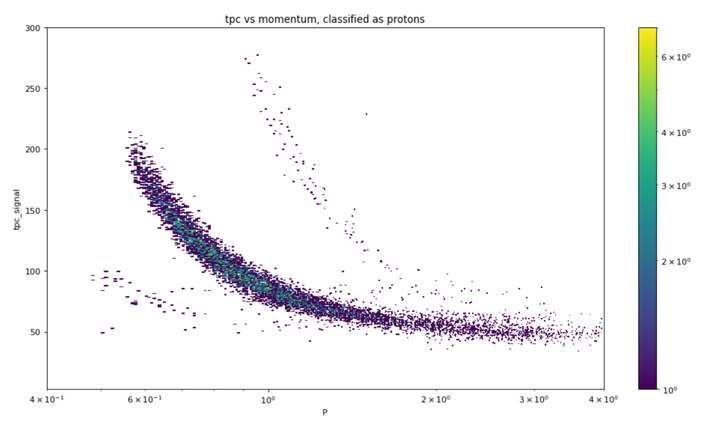

Figure 6:Particles classified by model with domain adaptation as protons, *TPC Signal* histogram.

Conclusion

- Domain adaptation techniques improve model accuracy on prodouction data.
- Further investigation on model accuracy measurments is needed.

Acknowledgements

This work has been supported by the Polish National Science Centre, grant number 2016/22/M/ST2/00176, the Ministry of Science and Higher Education, decision number DIR/WK/2016/2018/17-1, and by IDUB-POB-FWEiTE-1 project granted by Warsaw University of Technology under the program Excellence Initiative: Research University (ID-UB).