Machine learning techniques for optimisation of track selection criteria

Evgeny Andronov, Igor Altsybeev, Daria Prokhorova
Saint Petersburg State University

Introduction

One of the ways to improve the analysis of experimental data is to optimize track selection criteria. Using Monte Carlo simulations, one can train machine learning classifiers to separate correctly reconstructed primary tracks from secondary and fake tracks based on their features such as a number of clusters in TPCs, distance of closest approach to an interaction vertex etc.

In this contribution we performed the analysis based on the Monte Carlo simulations of inelastic proton-proton interactions within the NA61/SHINE experimental facility.[1]

NA61/SHINE is a fixed target experiment located at the CERN SPS. It has a complex geometry of tracking detectors resulting in a non-trivial behaviour of reconstruction efficiency in different kinematic acceptances.

Methods

Classifiers:

- Linear Discriminant
- Quadratic Discriminant
- AdaBoost[2]
- Neural network (Multi-layer perceptron with 100 neurons in a single hidden layer)

Implemented within scikit-learn[3]

```
lda = LinearDiscriminantAnalysis(solver="svd", store_covariance=True)
qda = QuadraticDiscriminantAnalysis(store_covariance=True)
ada = AdaBoostClassifier(DecisionTreeClassifier(max_depth=2),algorithm="SAMME.R",n_estimators=200)
nn = MLPClassifier( solver='adam',  learning_rate='adaptive', hidden_layer_sizes=(100,), alpha=0.0001,
activation='relu' )
```

Dataset:

- EPOS1.99 model[4], $p+p$ at p_{lab}=20 GeV/c
- 10 millions of events
- 70/30% split into 'train' and 'test' samples
- Full detector simulation of NA61/SHINE based on GEANT3

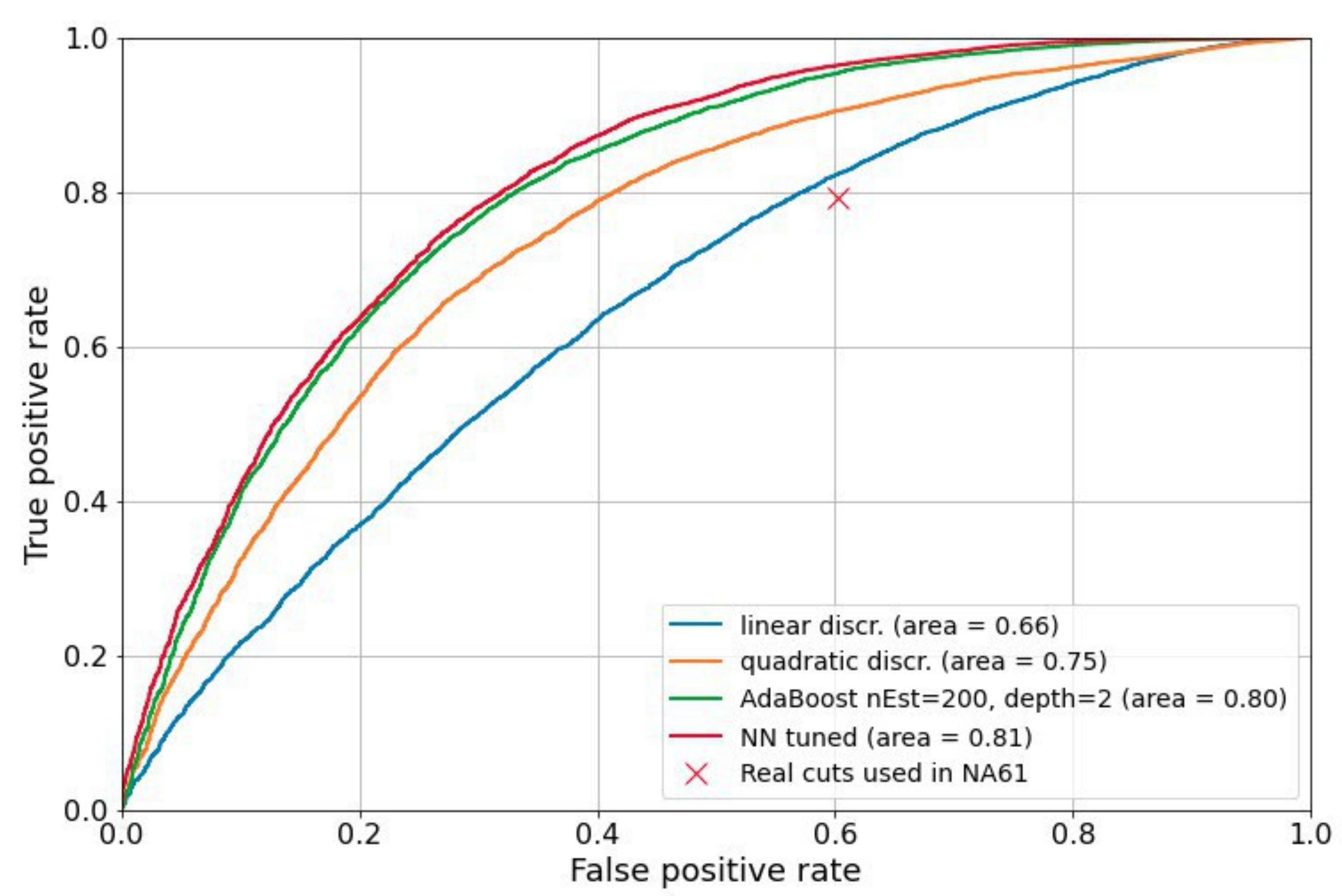
Labelling

At the first step all reconstructed tracks were labelled using a track matching procedure as proper primary tracks and others (fakes and secondaries). The tracks were matched to simulated ones using information about hits that original tracks left in TPC during GEANT simulation and reconstructed clusters.
For each reconstructed track a number of features corresponding both to the track and a given event was recorded.

Features:

- Momentum vector
- Distance of closest approach to the vertex
- Electric charge
- Number of clusters in TPCs
- Event multiplicity
- Vertex position

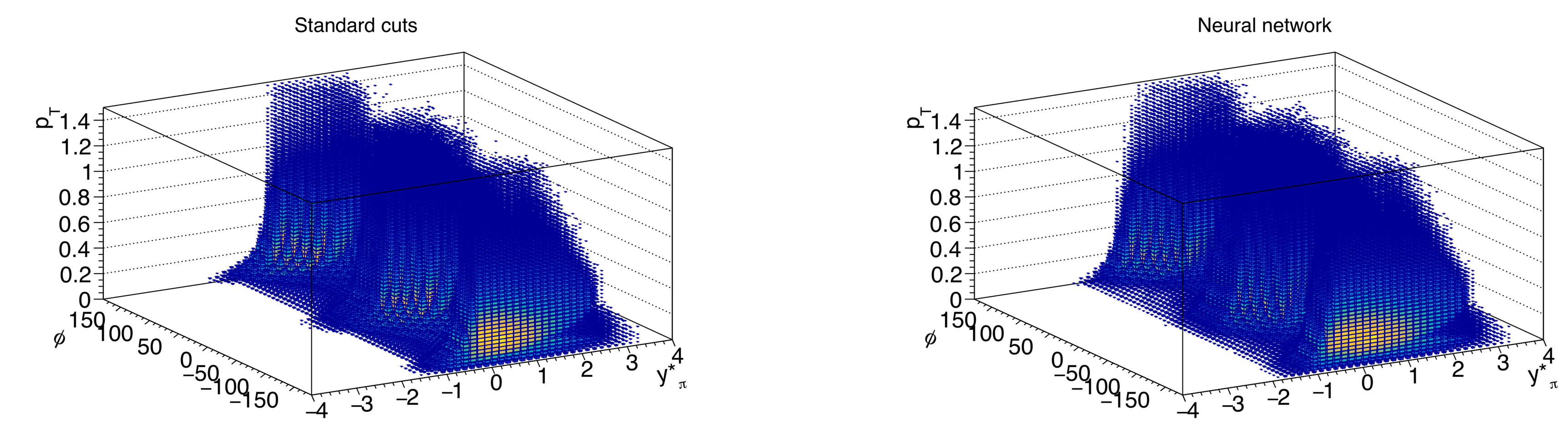
Training results (ROC curve)



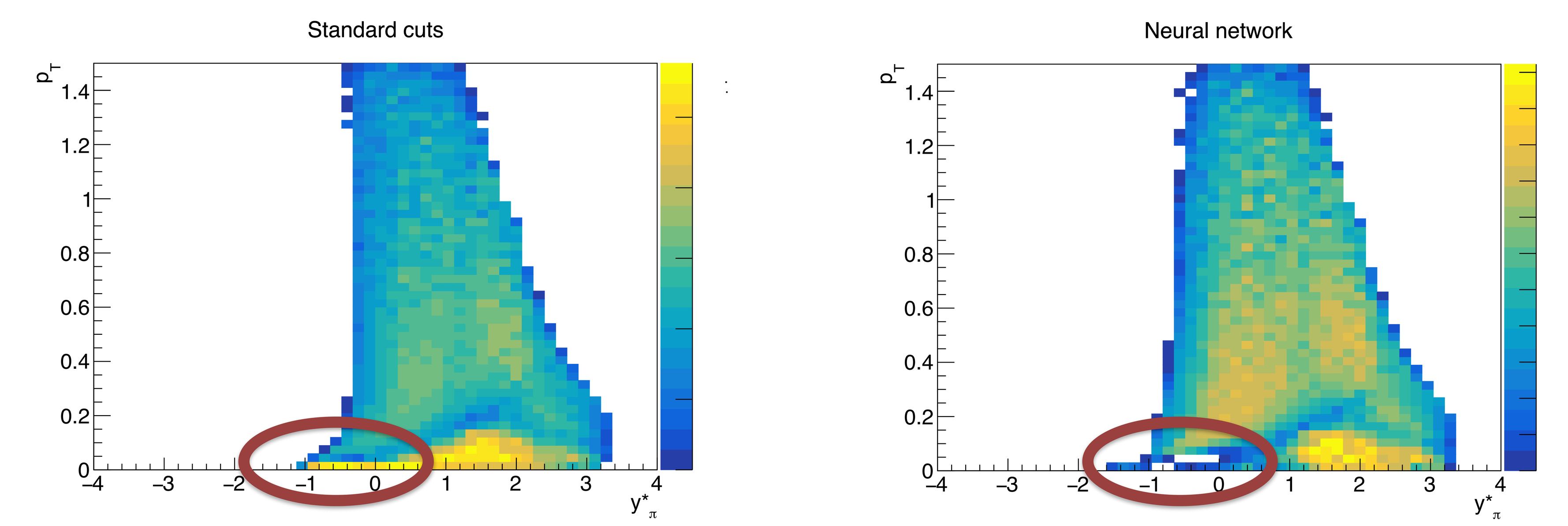
Application of ML classifiers allows us to significantly increase the selection efficiency of true primary tracks, as compared to the selection criteria currently used in experiment (red cross marker).

Acceptance maps

NA61/SHINE acceptance map[5] is a 3d histogram (p_T, y, phi) that was constructed by selecting kinematic regions where number of reconstructed tracks selected using standard criteria was close to the number of generated tracks.
One can repeat this procedure with track selection executed by a ML classifier. Populations of reconstructed tracks obtained with standard track selection and with the neural network are presented below.



Now we can divide these populations to the original one and construct two acceptance maps by selecting regions with high efficiency. In order to improve visual clarity, we project the obtained maps in phi direction.



This study showed that with application of neural nets one can optimise track selection. One can remove regions with a large proportion of non-primary tracks from the analyzed acceptance (e.g. y<0, low pT) and add areas (e.g. y<0, pT>0.2 GeV/c) by reducing the requirements for tracks in these regions while maintaining their high-quality reconstruction.

References

1. Abgrall N (NA61/SHINE Coll.) 2014 *JINST* **9** 06005.
2. Pedregosa F et al. 2011 *JMLR* **12** 2825.
3. Freund Y, Schapire R 1997 *JCSS* **55(1)** 119.
4. Werner K 2007 *Phys. Rev. Lett.* **98** 152301.
5. NA61/SHINE acceptance map https://edms.cern.ch/document/1549298/1

This work is supported by the Russian Science Foundation under grant 17-72-20045.
We thank to the support and help from all the members of the NA61/SHINE Collaboration.

e.v.andronov@spbu.ru