Machine learning techniques for optimisation of track selection criteria

Igor Altsybeev, Evgeny Andronov and Daria Prokhorova
Saint-Petersburg State University, Universitetskaya nab. 7/9, St. Petersburg, 199034, Russia
E-mail: e.v.andronov@spbu.ru

Abstract. Application of machine learning (ML) algorithms in high-energy physics is evolving rapidly. In particular, they could be used for the optimization of track selection criteria in the analysis of experimental data on hadronic collisions. Using Monte Carlo simulations, one can train ML 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 paper, we present the procedure of track selection optimization based on ML techniques and applied to EPOS1.99 simulations of proton-proton interactions obtained via Shine Offline Framework. With this approach, an increase of a fraction of the selected primary tracks and reduced contamination by the secondary tracks is obtained. In case of a complex geometry of an experimental facility like NA61/SHINE, improvement of track selection leads also to a widening of the kinematical acceptance.

1. Introduction
NA61/SHINE \[1\] is a fixed target experiment located at the CERN SPS. It has a complex geometry of tracking detectors resulting in a non-uniform dependence of reconstruction efficiency on considered kinematic acceptance. Typical experimental analysis in this case is performed in the limits of an acceptance map that defines kinematic regions with high track reconstruction efficiency. In turn, definition of efficiency takes into account track selection criteria that are unified for all kinematic regions. In order to optimize this step, one can loose the selection criteria and train machine learning classifiers on Monte Carlo simulated data to select properly reconstructed primary tracks.

In this contribution, we performed such an analysis using the Monte Carlo simulations of inelastic proton-proton interactions within the NA61/SHINE experimental facility.

2. Methods
A number of machine learning algorithms available through the scikit-learn module \[2\] was tested. The simplest algorithm used for this study is Linear Discriminant Analysis, which is a classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes’ rule. Next (similar) method utilized in this analysis is Quadratic Discriminant, where the decision boundary is quadratic. More sophisticated methods were adopted as well. The first is the boosting algorithm AdaBoost \[3\], a meta-estimator, which creates a multiple instances of Decision Tree classifier (with depth 2), where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on these
In order to evaluate the performance of the considered models, we used a dataset of inelastic proton-proton interaction at $p_{\text{lab}} = 20 \text{ GeV/c}$ generated with the EPOS1.99 model [4]. The total statistics of 10 million of events was split into train and test samples with a ratio 70 to 30%. The generated particles were transported through the NA61/SHINE experimental setup in GEANT3 and reconstructed with the Shine Offline Framework [5].

Before feeding this dataset to the models, a labelling of tracks was performed. Namely, for each reconstructed track the label is assigned, which tells whether this track corresponds to a proper primary track, to fake track or to a secondary track. The labelling relies on the matching procedure: the reconstructed 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 (components of momentum vector, electric charge, number of clusters in TPCs, distance of closest approach (DCA) to the primary vertex) and to a given event (multiplicity, vertex position) was recorded. A labelled dataset was stored in a form of a ROOT tree.

3. Results

Figure 1 (a) represents the performance of the considered algorithms in terms of ROC-curves, which show fractions of correctly reconstructed primary tracks (true positive rate or efficiency) as a function of false positive rate (i.e. when fake and secondary tracks are (incorrectly) marked by classifiers as good ones.) For comparison, results obtained with the standard NA61/SHINE track selection are denoted by a red cross marker. It is evident from this Figure, that one can increase a true positive rate significantly by applying of the trained classifiers, keeping the same value of a false positive rate as we have with the classic selection.

The figure 1 (b) shows the importance of the particular features for the trained AdaBoost classifier. It can be seen that the most relevant variables are $x$ and $y$ projections of the DCA and components of the track momentum vector $p$.

Thus, from the performed Monte Carlo study it is clear that the considered machine learning
classifiers allow us to significantly increase the selection efficiency of true primary tracks, as compared to the selection criteria currently used in experiment.

Although direct application of the trained classifiers to experimental data may be a too big step forward, one can apply them to a redefinition of the acceptance map. NA61/SHINE acceptance map \[6\] is a three-dimensional histogram \((p_T, y, \phi)\) that was constructed by selecting kinematic regions where a 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 in figure 2.

![Figure 2](image)

**Figure 2.** Three-dimensional populations of reconstructed tracks selected with the standard criteria (a) and by neural network (b).

Now we can divide these populations to the population of generated primary tracks and construct two acceptance maps by selecting regions with high efficiency (> 90%). In order to improve visual clarity, we project the obtained maps in \(\phi\) direction (see figure 3). In panel (b), one may see a depletion in the area near \(y = 0\) and low \(p_T\) that corresponds to better rejection of the fake tracks in the ML-based analysis. In the rest of the plot, there is an overall increase in the acceptance.

![Figure 3](image)

**Figure 3.** Acceptance maps projected along \(\phi\) axis based on reconstructed tracks selected with the standard criteria (a) and by neural network (b).
4. Summary
In this work, we showed that with application of machine learning algorithms one can optimise quality of track selection. In a particular case of $pp$ collisions in NA61/SHINE experiment, one can remove regions with a large proportion of non-primary tracks from the analyzed acceptance (e.g. $y < 0, \, p_T < 0.2 \text{ GeV}/c$) and add areas (e.g. $y < 0, \, p_T > 0.2 \text{ GeV}/c$) by reducing the requirements for tracks in these regions while maintaining their high-quality reconstruction. One of the possible next steps would be an application of the new acceptance map to analysis of experimental data. It can be tested whether this would lead to the decreased systematic uncertainties of the results.

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

Appendix
Classifiers used in this work were initialized in the scikit-learn framework as is given below:

```python
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' )
```

References
[1] Abgrall N (NA61/SHINE Collaboration) 2014 J. Instr. 9 06005
[2] Pedregosa F et al. 2011 J. Mach. Learn. Res. 12 2825
[3] Freund Y, Schapire R 1997 J. Comput. Syst. Sci. 55 119
[4] Werner K 2007 Phys. Rev. Lett. 98 152301
[5] Sipos R, Laszlo A, Marcinek A, Paul T, Szuba M, Unger M, Veberic D, Wyszynski O 2012 J. Phys. Conf. Ser. 396 022045
[6] NA61/SHINE acceptance map URL https://edms.cern.ch/document/1549298/1