CERN

TurboSP and the Topological Trigger

Author:
Vladislav Belavin

Supervisor:
Alex Pearce

August 25, 2017

Contents

Abstract

TurboSP was originally proposed as an alternative to Full stream in LHCb data flow. TurboSP is a data flow strategy which not only selects events that should be preserved, like in Full stream, but also provides selective persistence. This is achieved by saving candidates and subset of the reconstruction.

During this summer project we investigated the physics viability of using TurboSP with the topological lines and found out a possibility to reduce the number of kept tracks per event by two times while keeping a ratio of fully picked up interesting decay modes on $\sim 97\%$ level.

1 Introduction

LHCb is a forward arm spectrometer with pseudorapidity range $2 < \eta < 5$. LHCb was originally designed for the study of b-physics. The trigger system of the experiment consists of hardware trigger(level-0 or L0) and software trigger(HLT1&2). Level-0 trigger reduces the input rate of events from 40 MHz to 1 MHz. HLT1&2 decrease it even further to the rate of 12.5 kHz.

The data at LHCb is saved in two streams: Turbo stream at the rate of 2.5 kHz and Full stream at the rate of 10 kHz. At the HLT2 stage since RUN 2 the quality of online reconstruction is the same as the quality of offline reconstruction. So it is possible to conduct physical analysis during the data collection.

In Full stream the RAW event information that satisfies special selection criteria is saved on the disk. An average size of the event in Full stream is ~ 70 kB. Afterwards physics analysis of RAW events requires offline reconstruction.

In Turbo stream only information on candidates that passed selection criteria is saved. These trigger lines have exclusive nature and high efficiency for specified decay modes. An average size of the event in Turbo stream is ~ 15 kB and an event contains only reconstruction on candidates.

At this moment ≈ 600Mb of data per second is kept for further offline processing. In 2020 for LHC RUN 3 instantaneous luminosity will be increased by 5 times from 0.4 nb^{-1}/s to 2 nb^{-1}/s and in 2025 in the HL-LHC era to 4 nb^{-1}/s. Which means that the output bandwidth to offline storage would also drastically increase.

TurboSP is a data flow strategy which not only selects events that should be preserved, like in Full stream, but also provides selective persistence, i.e. saves candidates and subset of the reconstruction. Such a design of TurboSP provides physicists with more information for analysis than Turbo stream while saving disk space and reducing output rate.

In this project we are investigating the physics viability of using TurboSP with the topological lines and multivariate classifier.

2 TurboSP and Topological Lines

2.1 Topological Lines

The topological trigger is at the core of Full stream. Topological lines in the trigger exploit general properties of B decays like high transverse momentum of daughters, large flight distance due to the long lifetime of B meson and avoid tight cuts on the quality of the vertices and mass of B candidate to remain exclusive with respect to all modes.

These topological lines have a range of good properties[1]:

- high efficiency for any B decay, due to the inclusive nature of the trigger lines;
- an excellent timing performance and background rejection, due to the small number of trigger lines;
- excellent data-mining properties due to their inclusive nature;
- trigger redundancy for golden modes.

For simplicity, we would call a set of tracks that fired topological trigger(i.e. TOS with respect to the trigger) as a topological candidate. Topological candidates could not cover all decay tracks, moreover usually only a few of the decay tracks fire the topological trigger(fig. 1). For this reason, it is inefficient to save on the disk only topological candidates.

However, topological trigger and topological candidates could be used as a seed for further analysis and track selection.

2.2 TurboSP Modification

The core of proposed modification(fig. 2) is a combination of topological trigger and multivariate classifier.

The topological trigger is used as a seed for a classifier. Only those events that passed topological selection criteria would be considered. All long tracks that are TOS with respect to topological lines would be saved on the disk. Other long tracks are passed to the classifier and only tracks with a response higher than some threshold are kept.

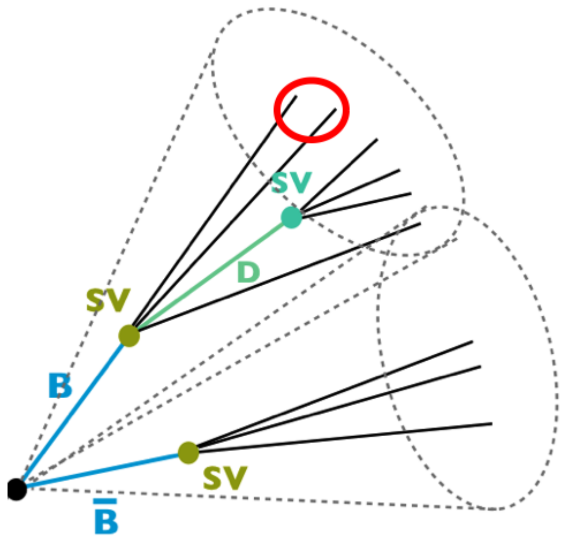


Figure 1: Example of decay topology. The ed circle highlights tracks that fired topological trigger

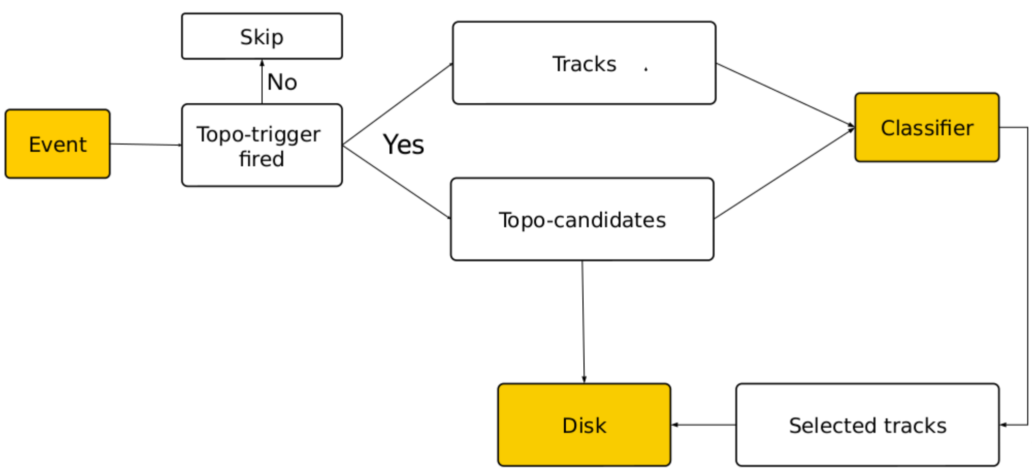


Figure 2: TurboSP proposed design

3 Data Analysis

In the analysis MC data the following decays were used:

- $B_s^0 \to \phi(1020) \to K^- K^+)(\phi(1020) \to K^- K^+)$
- $B^+ \to (D^0 \to K^+ \pi^-) K^+$
- $B^0 \to (D^0 \to K^+ \pi^-) K^+ \pi^-$
- $\Lambda_b^0 \to p^+ \pi^- \pi^+ \pi^-$
- $\Lambda_b^0 \to (D^0 \to K^- \pi +) p^+ K^-$
- $B_s^0 \to (K_0^*(892) \to K^+ \pi^-)(\bar{K}_0^*(892) \to K^- \pi^+)$
- $B_0 \to (D^+ \to K^- \pi^+ \pi^+) \pi^-$

Half of these decays represent charm physics and the other half represent charmless physics. Also, all of them slightly differ in topology so that the classifier would be able to generalize.

3.1 Preprocessing and Feature Extraction

For data preprocessing and feature extraction DaVinci script was written. It works with both BKQuery data and local files and takes as input a configuration file in the following format:

```
[decay]
name: Bu_D0K,Kpi=DecProdCut #short decay description
evttype: 12163011 #event type code
descriptor: [B+ => (D~0 => K+ pi-) K+]CC #please, find documentation
#for LoKi Decay Finder syntax in the references

[data]
year: 2012 #date of MC simulation
conddb: sim-20160321-2-vc-md100 #tags of MC conditions
dddb: dddb-20150928
bkquery: /MC/2012/Beam4000GeV-2012-MagDown-Nu2.5-Pythia8/Sim09b/
Trig0x409f0045/Reco14c/Stripping21NoPrescalingFlagged/
12163011/ALLSTREAMS.DST #BKQuery link to MC data
```

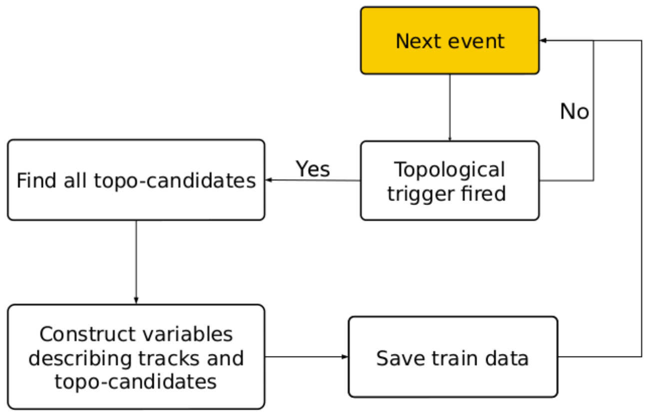


Figure 3: Data preprocessing

As an output it returns a serialized Pandas DataFrame with a set of variables (Appendix A) for each track per topological candidate. Each track has a label 1 if it is associated with an interesting decay in MC relation table and a label 0 in all other cases.

3.2 Model Training

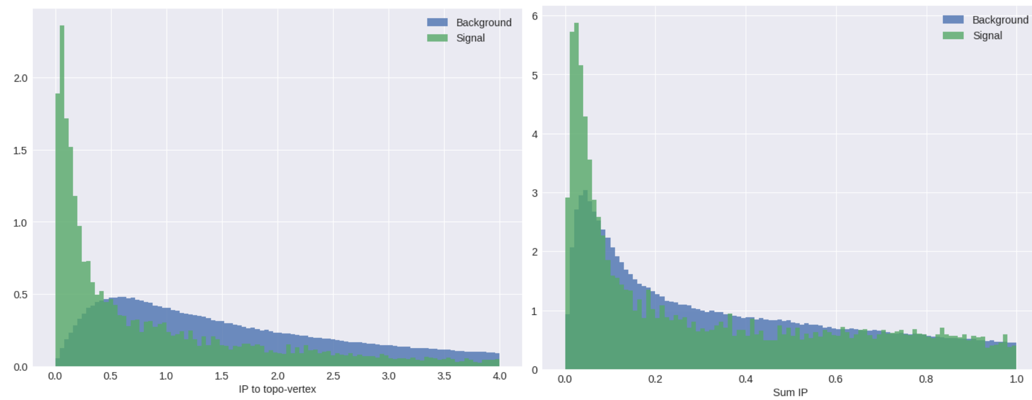
For the classification LightGBM [2] (realization of Gradient Boosted Decision Trees) was used. LightGBM has a variety of good features: serialization of trained model, support of CPU and GPU parallelization, speed-up by 2-10 times in comparison with XGBoost at both training and evaluation steps.

At Fig. 4a, 4b, 5a one can see distribution of several variables for signal and background tracks. The difference in distributions quite noticeable. Fig. 5b shows classifier response for signal and background.

To evaluate a performance of the classifier two metrics were proposed.

The first metric is ROC-AUC. ROC-AUC is the area under ROC-curve, which is also known in physics as a probability-probability plot. Another description of ROC-AUC is the proportion of correctly ordered pairs and can be estimated this way:

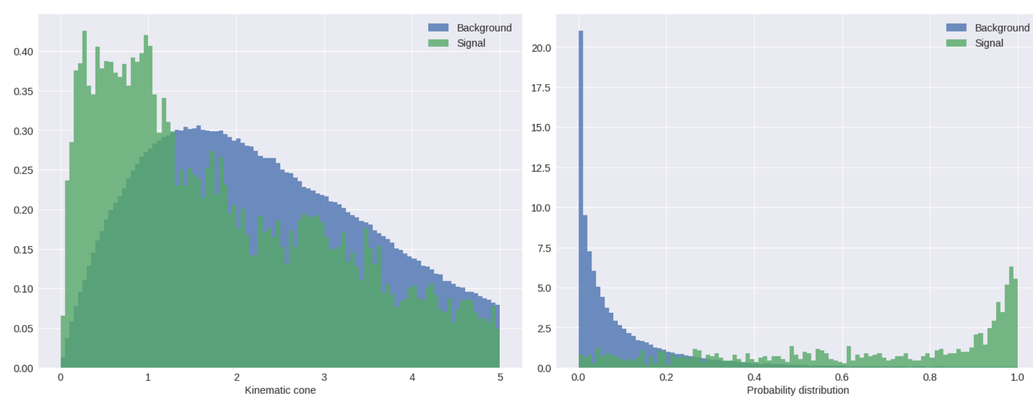
$$\mathrm{ROC} - \mathrm{AUC} = \frac{\sum\limits_{ij} I[y_i < y_j] I[p_i < p_j]}{\sum\limits_{ij} I[y_i < y_j]}$$



(a) IP to fitted vertex of topological candidate

(b) Sum of IP to fitted vertex of track and topological candidate system

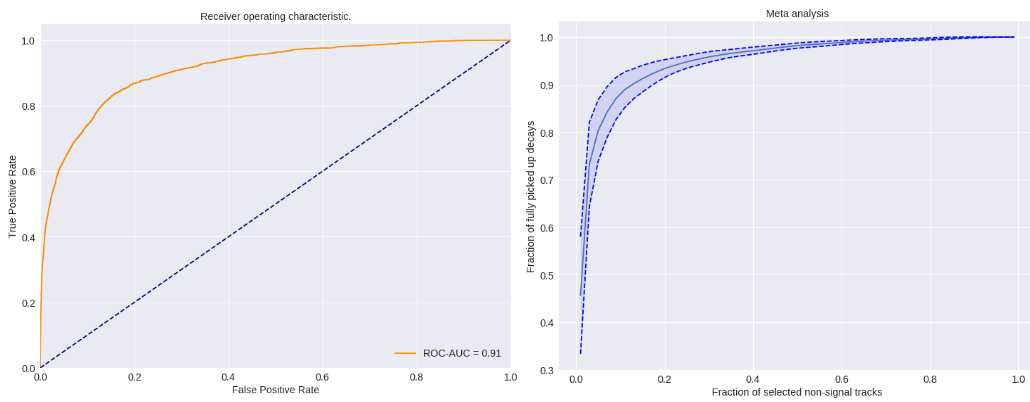
Figure 4: Distributions



(a) Kinematic cone

(b) Predictions distribution

Figure 5: Distributions



(a) ROC-curve (b) Trained on one. Quality: $94 \pm 3.5\%$

Figure 6: Evaluation

where p_i – is the probability of the track to be a signal, y_i – ground truth and I – indicator function. This metric is stable to small deviations in distributions and can be used in cases of unbalanced datasets. In the best case ROC-AUC=1, which implies perfect classification and in the worst case ROC-AUC=0.5, which means that classifier assigns labels randomly and uniformly. In Fig. 6a, one can see ROC-curve with ROC-AUC=0.91.

Another proposed metric is a curve where the x-axis shows a fraction of selected non-signal tracks and the y-axis shows a fraction of decays for which all available decay tracks are picked up, which could be interpreted as the efficiency of selection.

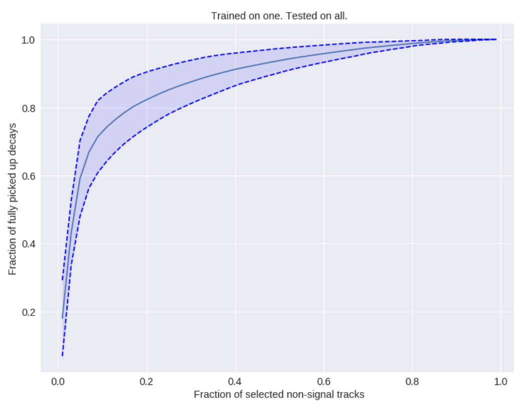
For instance, we are looking at the fraction of fully picked up decays while keeping only 50% of non-signal tracks in selection.

To understand the generalization error of the classifier several types of cross-validation have been conducted.

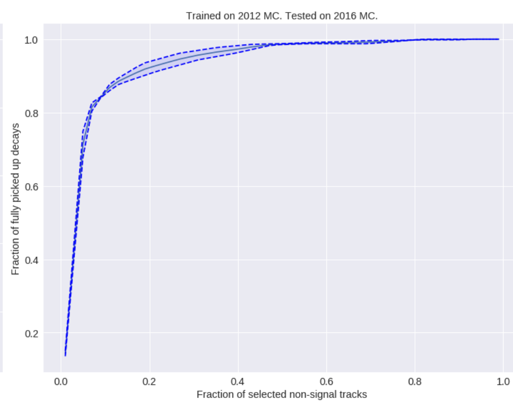
Firstly, the classifier is trained on one decay mode and tested on all other decay modes in order to estimate inclusiveness of selection(fig. 6b). An achieved quality of $94 \pm 3.5\%$ states that the classifier has good generalization properties.

Secondly, the classifier is trained on all decays but one on which it is evaluated. This is done in order to check if the quality decreases with a more general(trained on many decay modes) classifier(Fig. 7a). However, increase in the number of decay modes only enhance performance of selection.

Thirdly, the classifier is trained on all decay modes of 2012 MC data and evaluated on 2016 MC data in order to check the classifier robustness to a different condition(Fig. 7b). In 2012 the energy of the beam was 8 TeV and in 2016 the energy of the beam was 13 TeV. It can be seen that the classifier is unaffected by the change of the beam energy.



(a) Trained on all but one.
Quality: $96.4 \pm 0.6\%$



(b) Trained on 2012, evaluated on 2016.
Quality: $98.6 \pm 0.1\%$

Figure 7: Evaluation

4 Conclusion

The multivariate classifier shows good background rejection rate while keeping efficiency at a comparably high level. For instance, we can decrease the number of background tracks by two times while keeping a fraction of fully picked up decays on $\sim 97\%$ level.

Further steps:

- test physics performance on real data;
- test physics performance on MC data for higher instantaneous luminosity;
- integrate classifier in trigger.

5 References

References

[1] M. Williams, V. Gligorov, C. Thomas, H. Dijkstra, J. Nardulli, and P. Spradlin, "The HLT2 Topological Lines," 2011.

[2] Microsoft, "LightGBM." `https://github.com/Microsoft/LightGBM`.

[3] "LoKi-based Decay Finders." `https://twiki.cern.ch/twiki/bin/view/LHCb/FAQ/LoKiNewDecayFinders`.

[4] "The LHCb Starterkit lessons." `https://lhcb.github.io`.

A Variables

Track variables:

- p, p_T of the track;
- χ_2/N_{dof} of the track fit;
- ghost probability for the track;
- probability of χ_2/N_{dof};
- μ, ϕ – pseudorapidity and angle of the track;

To extract variables that represent topological candidates, vertex(V_{topo}) and mother particle(P_{topo}) for tracks in topological candidate were fitted.

- μ, ϕ of P_{topo};
- cardinality(i.e. number of tracks in the topological candidate).

To extract variables that represent relations between track and topological candidate, vertex(V_{comb}) and mother particle(P_{comb}) for **additional track** and tracks in topological candidate were fitted. Associated PV is a closest primary vertex to all tracks in the topological candidate. IP stands for impact parameter.

- IP and IP_{χ_2} of tract to the associated PV;
- IP and IP_{χ_2} of track to the V_{topo};
- p_t of the P_{comb}
- mass and corrected mass($\sqrt{|p_t|^2 + M^2} + |p_t|$) of the P_{comb};
- χ_2 of the V_{comb};
- $\sum\limits_{\text{tracks} \in \text{topo}-\text{candidate}+\text{track}} IP$ and $\sum\limits_{\text{tracks} \in \text{topo}-\text{candidate}+\text{track}} IP_{\chi_2}$ to the V_{comb};
- $\sqrt{(\phi_{track} - \phi_{P_{topo}})^2 + (\mu_{track} - \mu_{P_{topo}})^2}$ – a measure of closeness for track and topological candidate.