Reconstruction of Long-Lived Particles in LHCb CERN Project by Data Analysis and Computational Intelligence Methods

Grzegorz Gołaszewski[1], Piotr Kulczycki[1,2], Tomasz Szumlak[1], Szymon Łukasik[1,2]

[1] Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Kraków, Poland
{ggolasz,kulpi,szumlak,slukasik}@agh.edu.pl
[2] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Abstract. LHCb at CERN, Geneva is a world-leading high energy physics experiment dedicated to searching for New Physics phenomena. The experiment is undergoing a major upgrade and will rely entirely on a flexible software trigger to process the data in real-time. In this paper a novel approach to reconstructing (detecting) long-lived particles using a new pattern matching procedure is presented. A large simulated data sample is applied to build an initial track pattern by an unsupervised approach. The pattern is then updated and verified by real collision data. As a performance index, the difference between density estimated by nonparametric methods using experimental streaming data and the one based on theoretical premises is used. Fuzzy clustering methods are applied for a pattern size reduction. A final decision is made in a real-time regime with rigorous time boundaries.

1 Introduction

Particle physics represents one of the hottest areas for the applications of Machine Learning research. Computational intelligence, understood as a group of methods reacting to the environment in new ways, making useful decisions by imitating intelligent organisms or social mechanisms is being used to tackle tasks difficult to deal with, using standard statistical apparatus [3]; see also [7, 6].

The LHCb (Large Hadron Collider beauty) [8] experiment is one of eight particle physics detector experiments collecting data at the Large Hadron Collider (LHC), CERN, Geneva. It has been collecting proton-proton collision data during Run 1 and Run 2 data taking periods from 2010 to 2012 and from 2015 to 2018 respectively. Its primary physics mission is to search for New Physics and provide precise measurement of the charge-parity violation in the heavy quark sector [12]. Thanks to its extraordinary tracking system, in time, the physics programme has been extended and the detector became a general purpose in forward direction with a high performance high level trigger system [1]. As the experiment is currently undergoing a major modernisation increasing its instantaneous luminosity (and for the preparation for Run 3 that will start in 2021)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_25

the entire tracking system must be exchanged [11] and new readout electronics capable of on-detector raw data processing must be designed [10]. Most of the physics analyses in LHCb use, so called long tracks, i.e. trajectories of particles traversing the whole active volume of the LHCb tracking system.

The goal of this contribution is to propose a new algorithm, based on principles of data mining, for detecting and reconstructing so called long-lived particles in the revised detector scheme under new, more strict time and accuracy constraints. The problem of detecting such particles will be treated here as the instance of the classification task, that is assigning a tested element $\tilde{x}$ to one of the designated classes with known sets of representative elements. The reconstruction of long-lived particles for each analyzed track classifier should be assigned one of the label: 1 if the track is associated with the long-lived particle, and 0, if otherwise. The proposed algorithm relies on non-parametric construction of training patterns representing tracks of long-lived particles, with online adaptation of this training set. In addition to that, the algorithm uses fuzzy clustering for obtaining repeatedly a reduced representation of the training set.

2 Long-Lived Particles in LHCb

The most important, from the point of view of this paper, sub-systems of the upgraded spectrometer are those comprising the tracking system: vertex detector (VELO), upstream tracker (UT) and scintillating fibre tracker (FT); see Fig. 1. The VELO is a pixel detector located around the proton-proton interaction point and dedicated to precise measurement of charged particle trajectories. The UT detector provides additional measurements upstream of the magnetic field and is capable of reconstructing up to two space points. Finally, the FT is designed to sample trajectories after the magnet and provides up to three space points. The overall design goals of this system are to provide track reconstruction with efficiency greater than 95% and average momentum resolution of 0.5% for charged particles with momenta $p < 20$ GeV. Most of the physics analyses in LHCb use long tracks, which are trajectories of particles traversing the whole active volume of the LHCb tracking system.

The tracks reconstructed in the LHCb detector are divided into types depending on the sub-detectors in which they are observed. VELO tracks are defined as those which have hits only in the VELO detector. Upstream tracks are defined as those which have measurements only in the VELO and UT detectors. Upstream tracks are also referred to as Velo-UT tracks. FT tracks are defined as those which have hits only in the FT stations. Downstream tracks, which are the subject of this paper, must have hits both in the UT and FT trackers. Particles that are reconstructed as long tracks must traverse all of the tracking detectors. These tracks provide the best momentum resolution.

Downstream tracks, that are the signature of long-lived particles, are currently being reconstructed at LHCb using a state-of-art *PatLongLivedTracking* algorithms. The input of the algorithm constitutes a track segment container and hits reconstructed in the UT detector. The FT tracks are prepared by a standalone code that is not a part of the long-lived tracks reconstruction. In order to

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77961-0_25

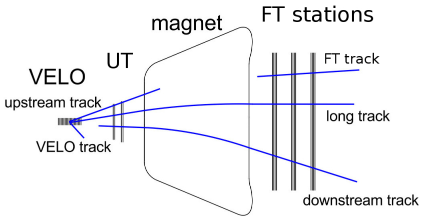

Fig. 1. A simplified diagram of the LHCb tracker system

increase purity of the input FT tracks they are processed by a multivariate classifier (binned Boosted Decision Tree) to discard bad candidates, e.g. segments that do not represent a real particle. Selected candidates are then propagated through the magnetic field to the UT tracker, assuming that they originate at the proton-proton interaction point. The UT detector comprises four layers of silicon micro-strip detectors divided into two stations UTa and UTb. Each station is capable of providing one space point (for each reconstructed trajectory) by probing x and y coordinates. The latter is measured by tilted by 5° and -5° layers UTau and UTbv also known as stereo layers, probing u and v coordinates respectively. The algorithm is searching for one hit in either of the x-probing layers which gives significant constraints on the flight trajectory of a particle. Next, an attempt is made to match a hit in the remaining x layer. Finally, hits are being searched for in the stereo layers. The hits are subsequently fitted with a parabola and χ^2 metric is calculated.

3 Proposed Approach

The algorithm being proposed here is divided into three separate phases. The initial stage includes data preparation and parameter adjustment. It is carried out before the experiment starts. This is followed by the on-line stage – the main phase from the point of view of the entire LHCb experiment. It is implemented in real time in the trigger circuit of the detector. Finally the calibration stage involves the modification of parameters and updating the on-line algorithm.

3.1 Initial stage

Generation of mass density training patterns. The algorithm starts with the creation of the f mass distribution density of detected long-lived particles for each of the two kinds of particles which are K^0_S i $\mathit{\Lambda}$. Due to the large number of factors affecting the measurements, the normal distribution is assumed in the form $f(m) = \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{-\frac{(m-\hat{m})^2}{2\hat{\sigma}^2}}$, where the value of the parameter $\hat{m}$ is established using physical premises. To determine the value of $\hat{\sigma}$, the Monte Carlo simulated events, prepared by the LHCb collaboration, is used (with the sample large enough for the estimator $\hat{\sigma}$ to be stable with assumed accuracy).
Construction of the pattern of pairs of long-lived particle tracks. The next step will be to create an n-element set W of unique reference pairs using the

Monte Carlo detector model. Such a set is created for each kind of particle and all described procedures is applied separately in regard of particle kind. Each pair ($w^i \in W, i = 1, 2, ..., n$) is described by the following parameters: particle mass m_i; coordinates (V_x^i, V_y^i, V_z^i) of the vertex of primary particle decay; the set of points $(x_{a,1}^i, v_1^i, u_1^i, x_{b,1}^i), (x_{a,2}^i, v_2^i, u_2^i, x_{b,2}^i)$ belonging to the subsequent hits on UT measurement planes, for the 1st and 2nd track from the pair, respectively; FT-seed state-vectors $\boldsymbol{s}_1^i = (x_1^i, y_1^i, t_{x,1}^i, t_{y,1}^i, \frac{q}{p}{}_1^i)$ and $\boldsymbol{s}_2^i = (x_2^i, y_2^i, t_{x,2}^i, t_{y,2}^i, \frac{q}{p}{}_2^i)$ associated with the appropriate tracks. The pattern elements generated in this way are then sorted relative to the coordinate $x_{1,1}$ (hit from the first UT measurement plane, and the first track of each pair) and divided into sectors of length resulting from measuring accuracy. This allows indexing of the pattern elements, which will significantly speed up the search for appropriate pairs at a later stage. Each pair is also initially assigned the fuzzy membership $\mu_i = 1$ in the pattern set [4]. The pattern prepared in this way will then be transferred to the LHCb trigger environment.

3.2 Online stage

The following section describes the procedure for finding pairs of long-lived particle traces from the data set. Such a set consists of hits from all UT detector measurement planes and FT-seed state vectors from T1, T2, T3 detectors created in earlier procedures. Due to the high frequency of intersection of proton beams and because this procedure is only one of many carried out during the experiment, it is necessary that the algorithm uses the least amount of both time and memory resources. It must also be constructed with a perspective of parallel implementation (in case of the execution comparing created tracks at this stage of the algorithm is lost in practice). It consists of the following steps.

1. Hits from the UTa_x plane are assigned to the respective sectors created in the initial stage, which allows them to be associated with pairs of tracks from the pattern, assigned to a given sector or directly adjacent sectors.
2. For each pair of reference traces (w_i) to which at least one hit has been assigned, all UT measurement planes are searched in order to find the corresponding hits from these planes and the first trace from the pair. The similarity criteria used result from technical measurement accuracy.
3. If at least 7 out of 8 possible hits on the UT detector were found, the best-matched elements are found in the set of FT-seeds according to the corresponding similarity criterion.
4. When the FT-seeds found in this way are sufficiently similar to the reference ones for both pair of tracks, hits from the UT detector with matched FT-seeds are marked as true tracks and forwarded for further analysis, which is beyond the scope of the algorithm presented.

3.3 Calibration stage

As part of the experiment, calibration procedures are carried out, improving the effectiveness of the procedure and adapting it to changing conditions. What's more, in the period between experiment breaks, in parallel to the on-line algorithm running, an update of the track pair patterns is launched and during the next break, an improved pattern is loaded into the system. Data for modifying the pattern are read from the buffer. Because the collected particles come from the very end of the reconstruction process, they are assigned a mass of m_i, along with all parameters that describe pattern pairs.

Adding new elements to the pattern. For all new w'_j elements, their similarity d to the most similar element from the W pattern is determined: $\Delta_j = \min\limits_{w_i \in W} \{d(w_i, w'_j)\}$, on the basis of which the decreasing function of membership of an element to a pattern is determined. If the element most similar w_i comes from the experiment, then their degrees of belonging become equal to $\mu_j = \mu_i$, while if w_i comes from the Monte Carlo model, the membership is

$$\mu_j = \begin{cases} 1 - 2\frac{\Delta_j^2}{{\epsilon_\mu}^2} & \text{, for } 0 \leq \Delta_j \leq \frac{\epsilon_\mu}{2} \\ 2\frac{(\epsilon_\mu - \Delta_j)^2}{{\epsilon_\mu}^2} & \text{, for } \frac{\epsilon_\mu}{2} < \Delta_j \leq \epsilon_\mu \\ 0 & \text{, for } \Delta_j > \epsilon_\mu \end{cases} , \tag{1}$$

where $\epsilon_\mu > 0$ describes the maximum similarity deciding element's acceptance.
Removing redundant elements. As the pattern increases in size, it is necessary to remove excess elements. For this purpose, the elements of the pattern are sorted according to the degree of the membership μ_i, and then $k < n$ elements with the lowest membership μ_i such that $\hat{f}(m_i) - f(m_i) > 0$ are removed from the pattern. In the above formula, f denotes the reference mass distribution created at the beginning of the procedure, while $\hat{f}$ describes nuclear mass distribution given in the form of the kernel estimator $\hat{f}(m) = \frac{1}{nh}\sum_{n=1}^{n} K(\frac{m-m_i}{h})$ where $K : \mathbb{R} \to [0, \infty)$ denotes the assumed kernel function and $h > 0$ means the smoothing parameter calculated by optimizing criteria [5].
Fuzzy pattern clustering. If the maximum acceptable size of the pattern is exceeded, while maintaining high values of μ, it is grouped by the fuzzy c-means clustering procedure [9]. Members of each of the K clusters are replaced by a representative (in the form of a centroid of a given cluster) with the weight $\psi_i = \sum_{j=1}^{n} c_{i,j} s_j$, where $c_{i,j}$ is a membership of the element j to cluster i. In addition to reducing the size of the pattern, this procedure also allows the gradual removal of Monte Carlo elements from the pattern, by replacing them with elements designated by means of experiments. The set of patterns for particular particle kinds is clustered separately.

4 Evaluation Criteria and Preliminary Results

The most important figures used to assess the performance of the downstream tracking algorithm are reconstruction efficiency and ghost fractions. Both are measured using simulated data samples by counting the number of correctly reconstructed tracks and compared to the number of tracks that would be reconstructed by a perfect algorithm. The latter tracks are called reconstructible. This corresponds to the concept of classification accuracy [2].

DOI: 10.1007/978-3-030-77961-0_25

The following definitions are used within the LHCb experiment. A particle is reconstructible as a downstream track if it is reconstructible as a FT track and has at least one hit in both UTa and UTb stations. A particle is reconstructible as a FT track if it has at least one x and one stereo hit associated to it in each of the three FT stations. A particle is considered reconstructed as a downstream track if at least 70% of the FT-station hits on a track are associated to it and the track has no more than one wrongly associated UT hit. Based on these definitions one can construct two types of efficiencies: the overall tracking efficiency $\epsilon_{rec,tot}$ and the efficiency related to the algorithm itself $\epsilon_{rec,alg}$. The former is useful for physics studies and represents the total reconstruction efficiency, i.e. corresponds to the number of downstream reconstructed tracks divided by the number of downstream reconstructible tracks. The latter also depends on the efficiency of the FT tracking and is calculated by dividing the number of downstream reconstructed and FT reconstructed tracks by the number of downstream reconstructible and FT reconstructed tracks. At the same time, it is worth mentioning that in our analyses electrons are discarded, since they undergo multiple scattering and more energy loss than other heavier particles. In addition to the efficiency functions ghost tracks must be also considered. Ghosts are a consequence of large particle occupancies and finite detector granularity. We define the fraction of the ghost tracks as R_{ghost} which corresponds to the number of ghost tracks related to the total number of downstream reconstructed tracks.

The presented algorithm has been preimplemented and integrated with the LHCb test environment. Tests were carried out on data generated in a Monte Carlo simulation and from real data from previous experiments. Because the actual data consists solely of measurements, it is not known which of them represents true particle tracks and there is no real possibility to fully reliably determine the algorithm efficiency. A set of training patterns was generated first. It consisted of about 250000 elements. A testing dataset, which was separately generated, consisted of the following particles: UT+T – 3815 particles, UT+T>GeV – 2327, UT+T_strange – 273, UT+T_strange>5GeV – 122, noVelo+UT+T_strange – 186, noVelo+UT+T_ strange>5GeV – 73 and 65621 particles of other types excluding noise. The tracking efficiency ($\epsilon_{rec,tot}$) was calculated with regards to the different particle types. The performance of the algorithm was demonstrated in Table 1. Taking into account the preliminary stage of the research it was found to be very promising. In all cases the majority of particles belonging to each of the groups were correctly identified.

Particle type	Tracking efficiency
UT+T	62.6
UT+T>GeV	75.6
UT+T_strange	58.8
UT+T_strange>5GeV	77.8
noVelo+UT+T_strange	54.5
noVelo+UT+T_strange>5GeV	66.7

Table 1. Tracking efficiency $\epsilon_{rec,tot}$ for tested particle types.

5 Final comments

This paper presents an algorithm for detecting traces of long-lived particles using elements of data analysis and computational intelligence. The aforementioned procedure is subject to numerous quality requirements as well as restrictions, especially the time needed for its implementation. This implies great possibilities to modify the concept during its implementation, depending on the partial results obtained. One of the planned research directions is the addition of a data validity mechanism known from data stream mining to the calibration stage, which would allow, in the long term, to remove pattern elements for which similar tracks are not found. In addition, this could improve the ability of the algorithm to adapt to variable physical parameters of the detector.

Acknowledgments

We acknowledge support from CERN and LHCb and from the national agency: MEiN and National Science Centre (Poland) UMO-2018/31/B /ST2/03998. The work was also supported by the Faculty of Physics and Applied Computer Science AGH UST statutory tasks within subsidy of MEiN.

References

1. Aaij, R., et al.: Tesla: An application for real-time data analysis in High Energy Physics. Computer Physics Communications **208**, 35 – 42 (2016)
2. Cady, F.: The Data Science Handbook. Wiley (2017)
3. Da, R.: Computational Intelligence in Complex Decision Making Systems. Atlantis Computational Intelligence Systems, Atlantis Press (2010)
4. Kacprzyk, J., Pedrycz, W. (eds.): Springer Handbook of Computational Intelligence. Springer (2015)
5. Kulczycki, P.: Kernel estimators for data analysis. In: Ram, M., Davim, J. (eds.) Advanced Mathematical Techniques in Engineering Sciences, pp. 177–202. CRC/Taylor & Francis (2018)
6. Kulczycki, P., Kacprzyk, J., Koczy, L., Mesiar, R., Wisniewski, R. (eds.): Information Technology, Systems Research, and Computational Physics. Springer (2020)
7. Kulczycki, P., Koczy, L., Mesiar, R., Kacprzyk, J. (eds.): Information Technology and Computational Physics. Springer (2017)
8. LHCb collaboration: LHCb detector performance. International Journal of Modern Physics A **30**(7) (2015)
9. Miyamoto, S., Ichihashi, H., Honda, K.: Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications. Studies in Fuzziness and Soft Computing, Springer (2008)
10. Steinkamp, O.: LHCb Upgrades. Journal of Physics: Conference Series **1271**(1) (2018)
11. Szumlak, T.: Events reconstruction at 30 MHz for the LHCb upgrade. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **936**(1) (2019)
12. Vecchi, S.: Overview of recent LHCb results. EPJ Web of Conferences **192**(24) (2018)