Towards automation of data quality system for CERN CMS experiment

Authors: Maxim Borisyak\textsuperscript{1,2}, Jean-Roch Vlimant\textsuperscript{3}, Andrey Ustyuzhanin\textsuperscript{1,2}, Maurizio Pierini\textsuperscript{4}, Maria Stenina\textsuperscript{5} and Dmitry Smolyakov\textsuperscript{2}

\textsuperscript{1} National Research University Higher School of Economics; \textsuperscript{2} Yandex School of Data Analysis; \textsuperscript{3} California Institute of Technology; \textsuperscript{4} CERN; \textsuperscript{5} Yandex

22\textsuperscript{nd} International Conference on Computing in High Energy and Nuclear Physics, San Francisco, October 10-14, 2016

Introduction

Traditionally, quality of the data at CERN CMS experiment is determined manually which requires tremendous amount of person power. In this work, we describe an approach for automated Data Quality system.

Data and feature extraction

CERN open portal data, 2010.
Over 2500 features were extracted. Each feature is defined by:
- \textit{stream}: minimal bias, muon or photon enriched;
- \textit{channel}: muons, photons, flows or calorimeter particles;
- \textit{quantile by particle momentum}: $\frac{5}{5}$, $\frac{4}{5}$, ... , $\frac{1}{5}$;
- \textit{physical property of particle}: $\eta$, $\phi$, $p_T$, $f_x$, $f_y$, $f_z$ or $m$;
- \textit{statistic within lumisection}: one of 5 percentiles, mean or standard deviation.

Additionally:
- total momentum of event;
- instant luminosity;
- number of particles in event.

Decision making

Possible labels:
- almost surely good (‘white zone’);
- almost surely contains an anomaly (‘black zone’);
- ambiguous (‘grey zone’).

\begin{tabular}{ccc}
  \textbf{automated decision} & expert decision & \textbf{automated decision} \\
  almost surely bad & ambiguous & almost surely good \\
\end{tabular}

Figure 1: Decision making schematics. Horizontal axis corresponds to classifier’s score.

Quality metrics

\begin{align*}
\text{Rejection Rate} &= \frac{\text{Rejected}}{\text{Total quantity of samples}} \rightarrow \text{min}; \\
\text{Pollution Rate} &= \frac{\text{False Positive}}{\text{True Positive} + \text{False Positive}} \leq \text{const}; \\
\text{Loss Rate} &= \frac{\text{False Negative}}{\text{True Positive} + \text{False Negative}} \leq \text{const}.
\end{align*}

Results

Remaining Manual work

Conclusions

- 20\% saved person power for Pollution and Loss rates 0.05\%;
- 80\% saved person power for Pollution and Loss rates 0.5%.

In addition, for data not labeled automatically system provides its estimates and hints for a possible source of anomalies which leads to overall improvement of data quality estimations speed and higher purity of collected data.