technische universität dortmund

SFB 876 Providing Information by Resource-Constrained Data Analysis

Active Simulation Data Mining

Mirko Bunse, Amal Saadallah, and Katharina Morik

TU Dortmund, AI Group, 44221 Dortmund, Germany

`{firstname.lastname}@tu-dortmund.de`

ECML-PKDD 2019

Interactive Adaptive Learning

Learning from Simulations

A simulation models how a system $\boldsymbol{s}$ evolves over time, given the parameters $\boldsymbol{\rho}$:

$$Sim_{\boldsymbol{\rho}}(\boldsymbol{s}_t, \Delta t) = \boldsymbol{s}_{t+\Delta t}$$

Multiple steps make a black-box data generator:

$$\underset{\text{parameters}}{\boldsymbol{\rho}} \longrightarrow \boxed{\text{simulation}} \longrightarrow \underset{\text{training data}}{(\boldsymbol{x}, y)}$$

- Simulated data replaces data from the real system
- Making predictions is faster than simulating

+

Active Sampling

Goal: Sample $\boldsymbol{\rho}$ for maximal learning efficiency

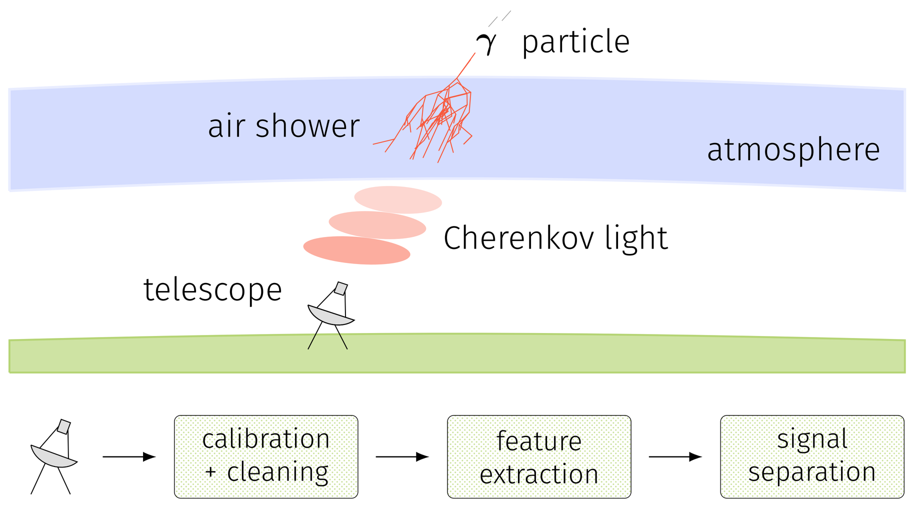
Do so sequentially, monitoring the trained model, and let either $\boldsymbol{x} \in \boldsymbol{\rho}$ or $y \in \boldsymbol{\rho}$, corresponding to the forward/backward distinction.

Cherenkov Astronomy

Labeled data only through simulation!

Simulation input: Label + auxiliary parameters (e.g. energy, direction, ...)

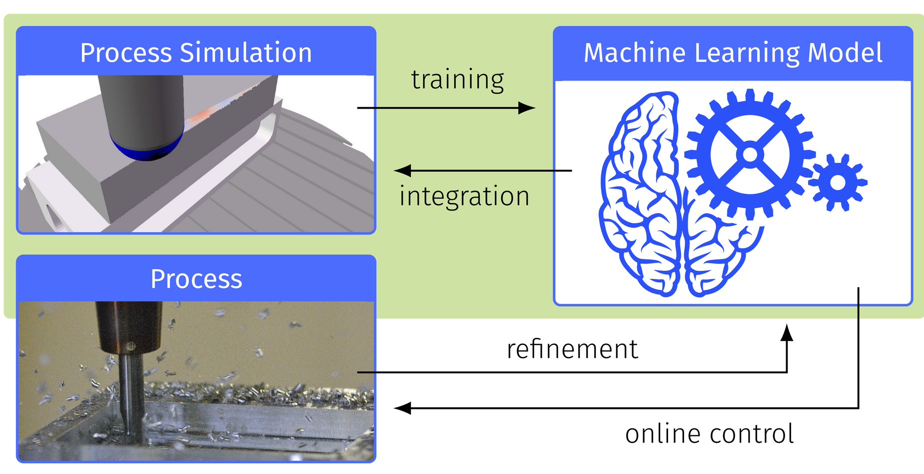
Simulation output: Synthetic observations



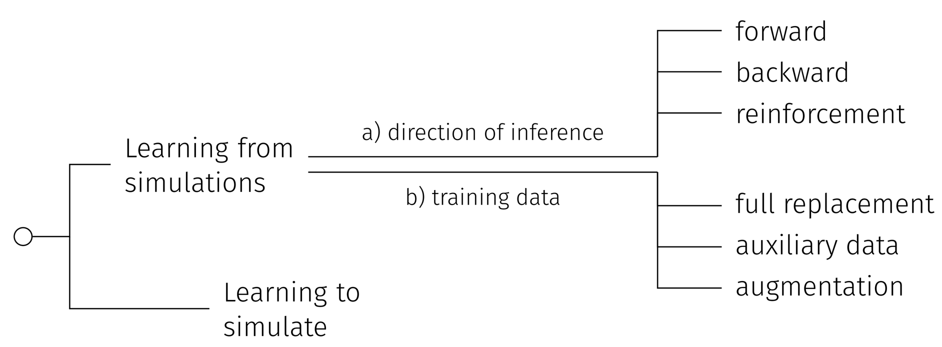
Milling

Simulation input: Process input parameters

Simulation output: Characteristics of the process

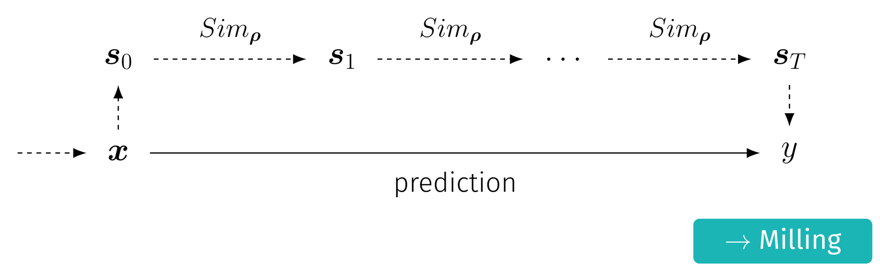


Other integrations of machine learning & simulation:

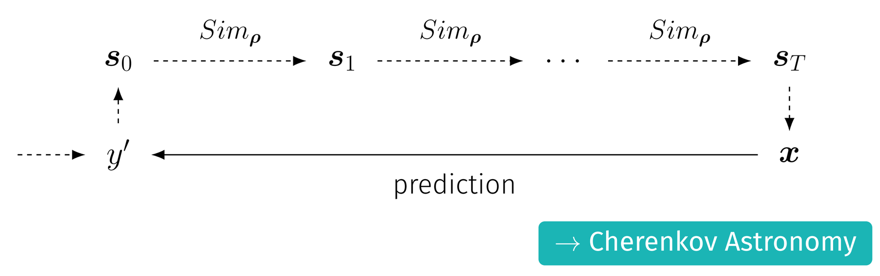


Forward vs. Backward

One may have to predict the **outcome** y...



...or the **cause** y' of an observation $\boldsymbol{x}$.



Example generation is either optimized by active learning (AL) or by active class selection (ACS), depending on the learning scenario.

In particular, the simulation candidates to score are either observations or labels:

$$u_{\text{AL}} : \mathcal{X} \rightarrow \mathbb{R}$$
$$u_{\text{ACS}} : \mathcal{Y} \rightarrow \mathbb{R}$$

Sampling of Parameters

The "pure" AL and ACS are *artificially* limited by neglecting the simulation parameters $\boldsymbol{\rho}$.

We expect that relevant data can be identified more easily by accounting for all parameters:

$$u : \mathcal{P} \rightarrow \mathbb{R}, \qquad \text{where} \begin{cases} \mathcal{X} \subseteq \mathcal{P} & \text{(AL)} \\ \mathcal{Y} \subseteq \mathcal{P} & \text{(ACS)} \end{cases}$$

Take this poster with you!

If all prints are gone,
please ask for a soft copy!

Conclusion & Outlook

We distinguish between the *forward* and *backward* learning scenario and propose simulation data mining as a use case for active sampling.

We identify open research issues:

Transfer Learning: The simulation may not exactly picture reality. Domain adaptation makes the two domains—simulation and reality—explicit.

Limits of ACS: In pure ACS, beating a merely random sampling is hard [5]. By accounting for all parameters $\boldsymbol{\rho}$, we hope to find whether this limitation comes from the narrow ACS task itself.

Data Imbalance in ACS: Between-class and within-class imbalances may harm ACS strategies.

Want to Collaborate?

`https://sfb876.tu-dortmund.de/simulation-data-mining`

References

[1] C. Bockermann, K. Brügge, J. Buss, A. Egorov, K. Morik, W. Rhode, and T. Ruhe. Online analysis of high-volume data streams in astroparticle physics. In *Proc. of the ECML-PKDD*, pages 100–115. Springer, 2015.

[2] A. Saadallah, F. Finkeldey, K. Morik, and P. Wiederkehr. Stability prediction in milling processes using a simulation-based machine learning approach. In *51st CIRP Conf. on Manufacturing Systems*. Elsevier, 2018.

[3] B. Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

[4] R. Lomasky, C. E. Brodley, M. Aernecke, D. Walt, and Mark A. Friedl. Active class selection. In *Proc. of the ECML*, pages 640–647. Springer, 2007.

[5] M. Bunse and K. Morik. What can we expect from active class selection? In *Lernen, Wissen, Daten, Analysen (LWDA) conf. proc.*, 2019. Accepted for publication.

Mirko Bunse

- Simulation Data Mining
- Deconvolution

...as being applied in Cherenkov astronomy (project SFB 876-C3).

Amal Saadallah

- Ensemble Methods
- Time Series Analyses

...using sensor data of production processes (project SFB 876-B3).

* Supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Data Analysis", projects C3 and B3.