Machine Learning techniques for signal selection

Artem Golovatiuk, Andrey Ustyuzhanin

University of Kyiv (Ukraine), Yandex (Russia)

Why do we need ML?

- Our goal:
  - Reducing the number of background events in potentially signal data
- Statistical approach:
  - limited by our physical understanding of the system
- Machine Learning approach:
  - can discover complex correlations between features, can be robust to insignificant variations in case of high input dimensions.

Algorithms performance metrics

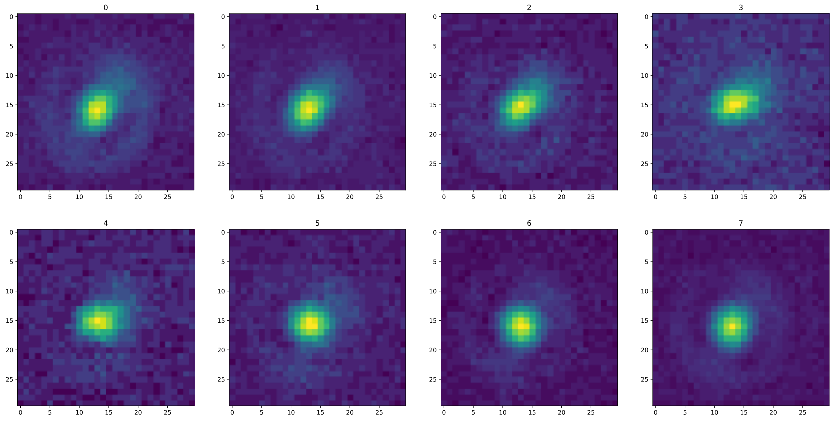
- Common metrics in ML: ROC–AUC score
- Physically motivated metrics: Precision
- $Precision\ = \frac{True\ Positive}{True\ Positive+False\ Positive}$
  - More background treated as a signal ⇒ lower $Precision$
- Use $Precision$ to check the performance of the final algorithm

Training data

- Signal: C 60keV, 80keV, 100keV
- Background: gamma, dust-fog
  - Lots of duplicates in the extracted images
- Current solution:
  - C 100keV signal vs LNGS exposed background
- Future plan:
  - Check performance for the signal-background pair-wise

Training data

- Gaussian fit parameters (7 polarizations):
  - x, y – cluster center coordinates
  - l_x, l_y – major and minor axes of an elliptical fit
  - φ – direction of the cluster
  - n_{px} – area of the ellipse in pixels
  - vol – volume of the cluster
    - 56 features in total
- Cluster images:
  - 8 polarizations
    for each sample



Tested approaches

- Convolutional Neural Networks:
  - Preliminary study by Sergey Shirobokov
  - Working directly with the cluster images
  - Shallow network, lots of room for improvement
  - Requires large computational power (e.g. GPU)

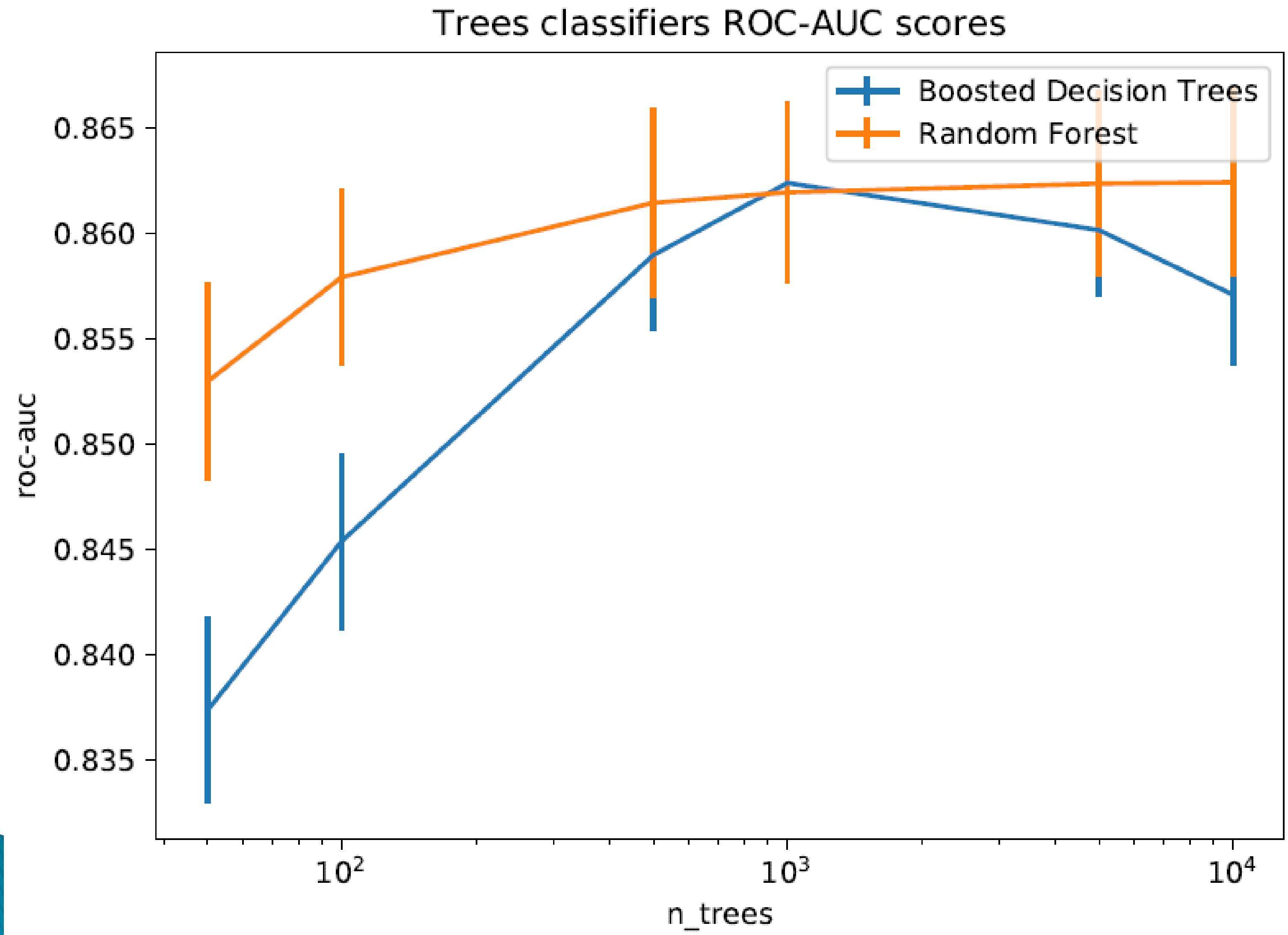
Tested approaches

- Boosted Decision Trees:
  - Composition of small decision trees, next one improves result of the previous one.
  - Limited possibility to parallelize
- Random Forest:
  - Composition of very deep trees, each one makes its own decision, result is the average of probabilities.
  - Highly parallelizable on CPUs
- Trees weakness:
  - Performance strongly depends on the features choice.

Preliminary results (ROC-AUC score)

- BDT
  - 0.86
  - 1000 trees
  - ~2 min
  - 10000 trees
  - ~23 min
  - ✓1 CPU
- RF
  - 0.86
  - 1000 trees
  - ~20 sec
  - 10000 trees
  - ~3.5 min
  - ✓20 CPUs
- CNN
  - 0.9
  - ~1000 epochs
  - ~30 hours
  - ✓1 GPU

Preliminary results (Trees)



Conclusions and further plans

- Potential of the approximate Gaussian fit approach is limited due to the information loss.
- Computationally cheap algorithms (e.g. composition of decision trees) can lower background contribution by order of magnitude, but have limited possibilities of improvement.
- Neural Networks can find complex correlations in the signal and background images directly, even when the difference is minimal. The broad variety of possible architecture types makes performance improvement limited only by the available computational power and amount of data.