AutoML Challenge: AutoML Framework Using Random Space Partitioning Optimizer

Jungtaek Kim[1], Jongheon Jeong[2], Seungjin Choi[1]

[1] Department of Computer Science and Engineering, POSTECH

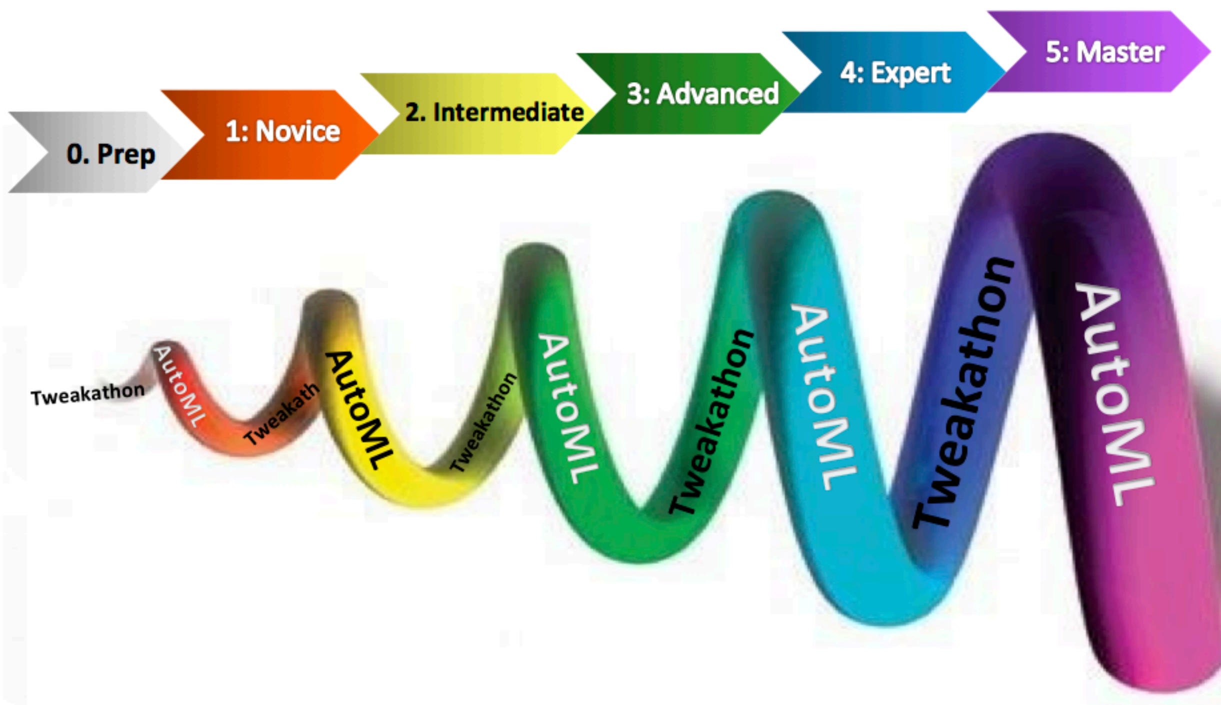
[2] exbrain Inc.

AutoML 2016 Workshop on ICML @ NYC

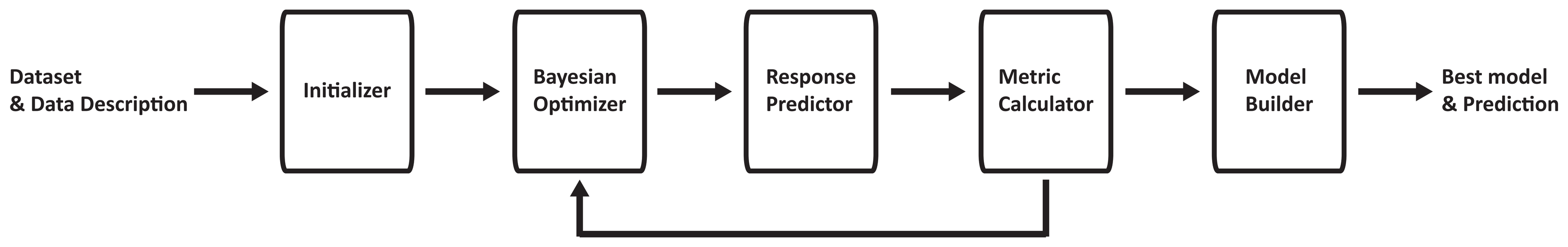
POSTECH XBRAIN

AutoML Challenge (Guyon *et al.*, 2015, 2016)

- Started in December 2014
- 5 rounds, excluding round 0
  - Binary classification / Multi-class classification / Multi-label classification / Regression
- 3 phases per each round
  - AutoML / Tweakathon / Final



Our Architecture



- Five components; meta-learning initializer, Bayesian optimizer, response predictor, metric calculator, and model builder
- Meta-learning initializer
  - Referred from auto-sklearn (Feurer *et al.*, 2015)
- Bayesian optimizer
  - Mondrian forests optimizer

Mondrian Forests Optimizer

Algorithm 1: Mondrian Forests Optimizer

Input: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where $\mathbf{x}_i \in \mathcal{ACS}$ and y_i is sampled from the performance measure, and Time budget $\mathcal{T}$
Output: $\mathbf{x}_{best} \in \mathcal{ACS}$

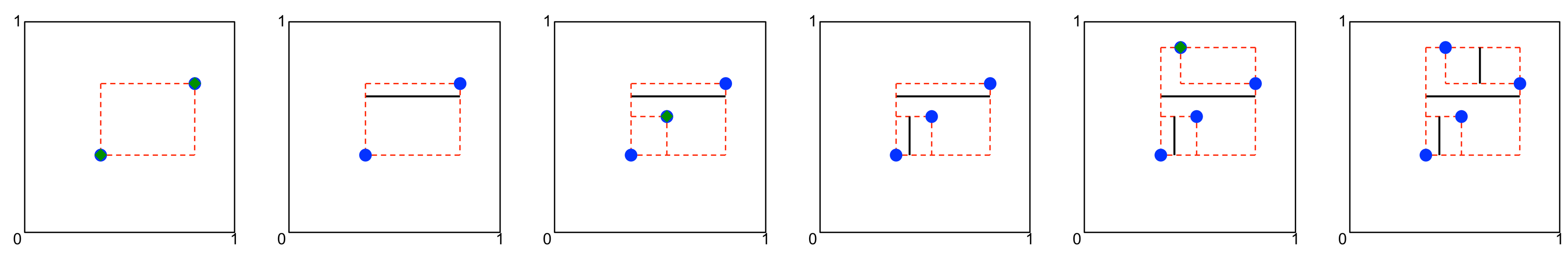
```
MF = None
for t < T do
    if MF == None then
        Build Mondrian forests, MF for D
    else
        Extend MF with {(x_new,j, y_new,j)}_{j=1}^K
    end
    Draw seed configurations ∈ ACS of local search for min_for_search times
    Search the neighbors of seed configurations and find the candidates, whose responses of the acquisition function
    are higher
    Merge the randomly sampled configurations ∈ ACS with the candidates queried from the acquisition function
    Update the best K configurations, {(x_new,j, y_new,j)}_{j=1}^K into D
end
return x_best ∈ ACS where x_best is the configuration which has the largest y_i of (x_i, y_i) ∈ D
```

- Random space partitioning optimizer
- Extended from Mondrian forests regression
- Handle all variables such as categorical and numerical variables
- Run on both Mondrian forests optimizer and actual response sampler in parallel

The Based System, *auto-sklearn* and Its Characteristics

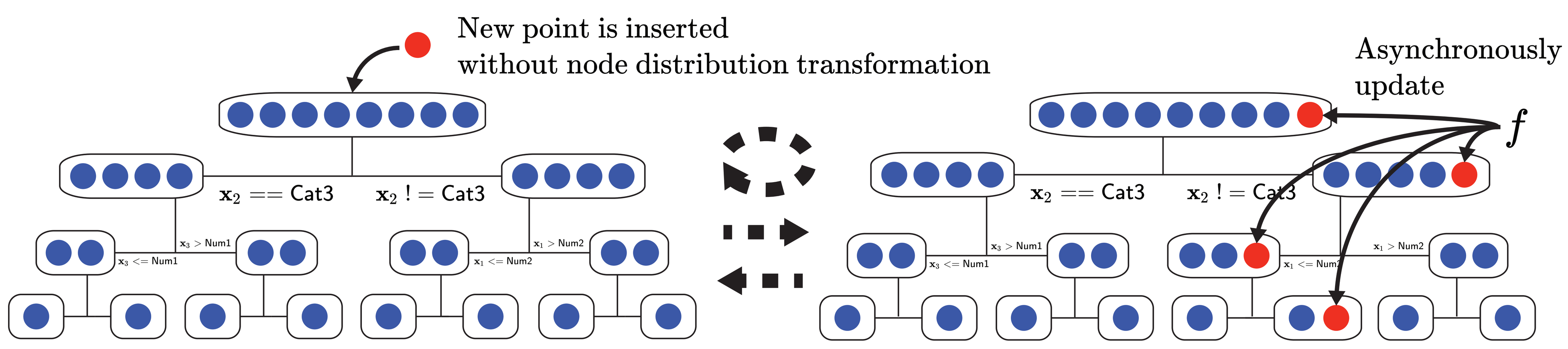
- Four components; meta-learning initializer, Bayesian optimizer, machine learning framework, and ensemble builder
- Based on scikit-learn library
- Optimized by SMAC (Hutter *et al.*, 2010)
  - Bayesian optimizer using random forests
  - Heuristic uncertainty estimation
  - Tree rebuilding is needed

Mondrian Forests Regression



- Introduced by Lakshminarayanan *et al.* (2016)
- An ensemble of probabilistic generalized k-d trees
- A restriction of a Mondrian process (Roy and Teh, 2008)
- A predictive label distribution of each tree is

$$p_{T_m}(y|\mathbf{x}_{\text{test}}, \mathcal{D}_{1:N}) = \sum_{j \in \text{path}(\text{leaf}(\mathbf{x}_{\text{test}}))} w_{mj} \mathcal{N}(y|\mu_{mj}, \sigma_{mj}^2)$$



AutoML Challenge Results

Final3 Team	Rank	Final4 Team	Rank	AutoML5 Team	Rank
aad_freiburg	1 (1.80)	aad_freiburg	1 (1.60)	aad_freiburg	1 (1.60)
djajetic	2 (2.00)	ideal.intel.analytics	2 (3.60)	djajetic	2 (2.60)
ideal.intel.analytics	3 (3.80)	abhishek4	3 (5.40)	**postech.mlg_exbrain**	3 (4.60)
asml.intel.com	3 (3.80)	**postech.mlg_exbrain**	4 (5.80)		
postech.mlg_exbrain	4 (5.40)				

Further Works and Conclusions

- Extend Mondrian forests optimizer to more straightforward assumption of Mondrian processes.
- Compare our system on a single machine and multiple machines
- Since AutoML is an online and sequential problem, Mondrian forests optimizer is proper to solve this problem.

Our System on GitHub

- https://github.com/postech-mlg-exbrain/AutoML-Challenge

References

- M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In Advances in Neural Information Processing Systems (NIPS), volume 28, 2015.
- I. Guyon, I. Chaabane, H. J. Escalante, S. Escalera, D. Jajetic, J. R. Lloyd, N. Macía, B. Ray, L. Romaszko, M. Sebag, A. Statnikov, S. Treguer, and E. Viegas. A brief review of the ChaLearn AutoML challenge. In Proceedings of AutoML 2016 Workshop on the International Conference on Machine Learning (ICML), 2016.
- I. Guyon, K. Bennett, G. Cawley, H. J. Escalante, S. Escalera, T. K. Ho, N. Macia, B. Ray, M. Saeed, A. Statnikov, et al. Design of the 2015 ChaLearn AutoML challenge. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), pages 1–8, 2015.
- F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration (extended version). Technical Report 10-TR-SMAC, UBC, 2010.
- B. Lakshminarayanan, D. M. Roy, and Y. W. Teh. Mondrian forests for large-scale regression when uncertainty matters. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2016.
- D. M. Roy and Y. W. Teh. The Mondrian process. In Advances in Neural Information Processing Systems (NIPS), volume 21, 2008.