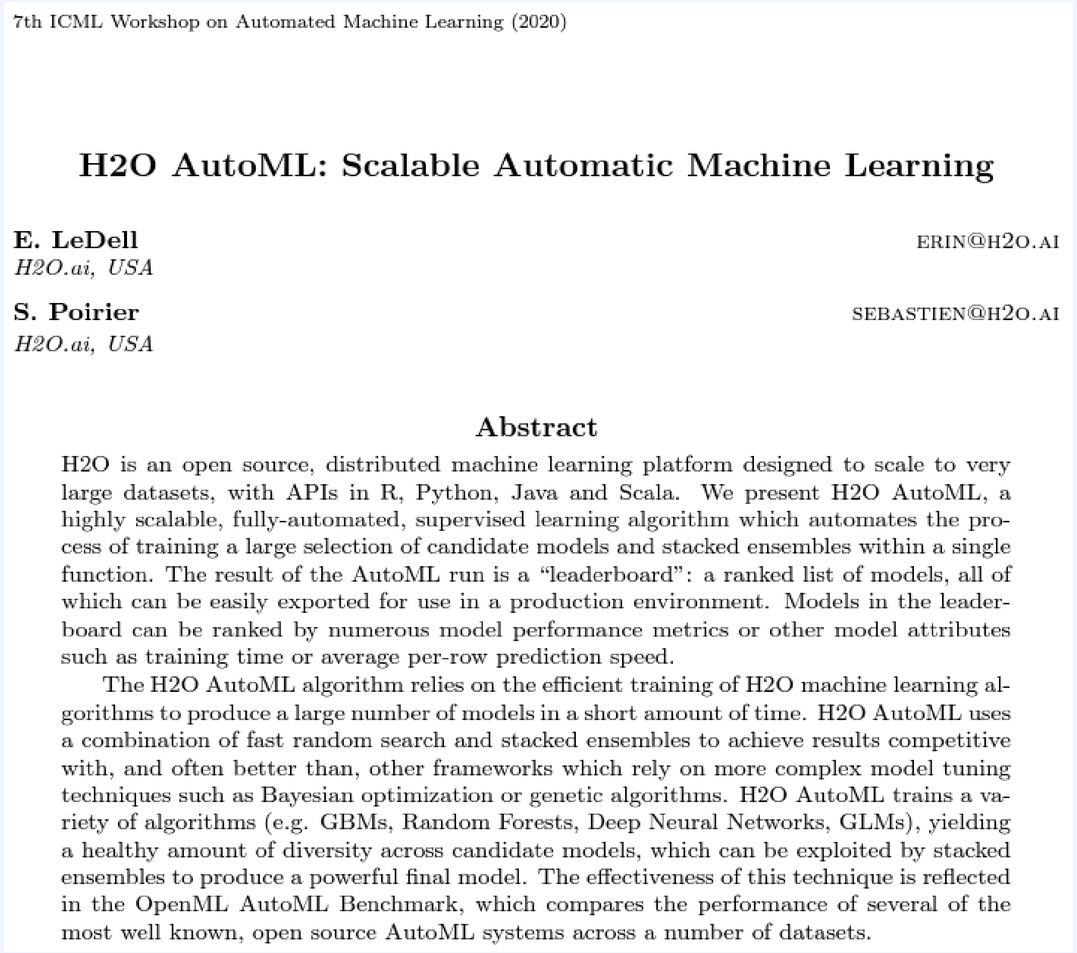
2. H2O

1 이론적 배경

관련 논문

H2O를 소개한 논문을 참고하여 H2O를 구성하는 이론적 배경에 대해 알아보겠습니다.



7th ICML Workshop on Automated Machine Learning (2020)

H2O AutoML: Scalable Automatic Machine Learning

E. LeDell ERIN@H2O.AI
H2O.ai, USA

S. Poirier SEBASTIEN@H2O.AI
H2O.ai, USA

Abstract

H2O is an open source, distributed machine learning platform designed to scale to very large datasets, with APIs in R, Python, Java and Scala. We present H2O AutoML, a highly scalable, fully-automated, supervised learning algorithm which automates the process of training a large selection of candidate models and stacked ensembles within a single function. The result of the AutoML run is a "leaderboard": a ranked list of models, all of which can be easily exported for use in a production environment. Models in the leaderboard can be ranked by numerous model performance metrics or other model attributes such as training time or average per-row prediction speed.

The H2O AutoML algorithm relies on the efficient training of H2O machine learning algorithms to produce a large number of models in a short amount of time. H2O AutoML uses a combination of fast random search and stacked ensembles to achieve results competitive with, and often better than, other frameworks which rely on more complex model tuning techniques such as Bayesian optimization or genetic algorithms. H2O AutoML trains a variety of algorithms (e.g. GBMs, Random Forests, Deep Neural Networks, GLMs), yielding a healthy amount of diversity across candidate models, which can be exploited by stacked ensembles to produce a powerful final model. The effectiveness of this technique is reflected in the OpenML AutoML Benchmark, which compares the performance of several of the most well known, open source AutoML systems across a number of datasets.

LeDell, E., & Poirier, S. (2020, July).
H2o automl: Scalable automatic machine learning.
In Proceedings of the AutoML Workshop at ICML (Vol. 2020).

탐색 공간 정의

H2O의 탐색 공간은 전처리와 모델 및 하이퍼파라미터로 구분할 수 있으며, 실험과 경험을 통해 정의합니다.

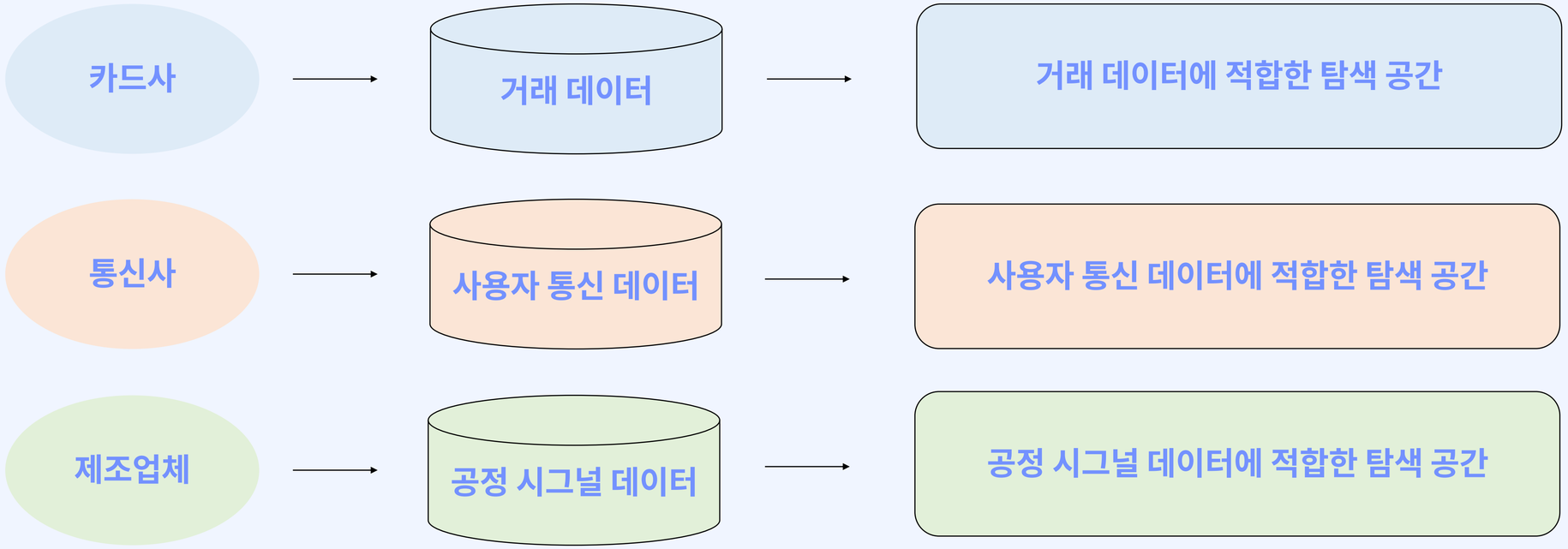
모델	그래디언트 부스팅 머신 (GBM)	XGBoost	랜덤 포레스트	심층 신경망	일반 선형 모델
데이터 전처리	결측치 추정	정규화	더미화	워드 임베딩	• • •

랜덤 포레스트의 하이퍼파라미터는 하나로 고정되는 등 탐색 공간이 매우 작음

→ 다양한 값을 탐색하는 데서 오는 성능 향상보다는 빠르고 신뢰성 있는 모델을 학습하는 것에 초점을 뒀기 때문

H2O의 탐색 공간으로부터 알 수 있는 점

H2O AutoML처럼 실험과 경험으로 탐색 공간을 적절히 정의하는 것이 머신러닝 자동화 시스템 개발에서 가장 중요하다고 생각합니다.



데이터 유형에 따라 적합한 탐색 공간이 다르므로, 커스터마이징된 머신러닝 자동화 시스템 개발이 필요함

최적화 알고리즘과 앙상블 모델

가장 단순한 방법인 랜덤 서치를 사용하여 모델 선택과 하이퍼 파라미터 튜닝 문제를 해결함

- 실험과 경험을 통해서 탐색 공간을 굉장히 좁게 설정했기 때문에 가능함
- XGBoost를 가장 먼저 탐색하고 다음으로 그래디언트 부스팅 머신을 탐색하는 등 탐색 순서와 각 모델을 탐색하는 시간의 비율도 정해져 있음

모델을 모두 학습한 뒤 두 개의 스태킹 앙상블 모델을 추가로 학습함

- 첫 번째 스태킹 앙상블 모델인 All Models는 학습한 전체 모델을 사용해 학습함
- 두 번째 스태킹 앙상블 모델인 Best of Family는 각 알고리즘(예: XGBoost, 심층 신경망 등)마다 최고 성능의 모델만 사용해 학습함
- H2O AutoML 개발자에 따르면, 일반적으로 이 두 앙상블 모델이 개별 모델보다 우수한 결과를 보인다고함

2. H2O

2 자동화 실습

패키지 설치

H2O AutoML은 다음과 같이 설치할 수 있습니다.

pip을 사용하는 경우

```
$ pip install -f http://h2o-release.s3.amazonaws.com/h2o/latest_stable_Py.html h2o
```

아나콘다를 사용하는 경우

```
$ conda install -c h2oai h2o
```

초기화

init 메서드를 사용하면 h2o를 초기화할 수 있으며, max_log_file_size, max_mem_size, min_mem_size 등의 인자로 분석 환경을 설정할 수 있습니다.

초기화

```
1  import h2o
2  h2o.init()
```

Checking whether there is an H2O instance running at http://localhost:54321 . connected.

H2O_cluster_uptime:	1 day 12 hours 52 mins
H2O_cluster_timezone:	Asia/Seoul
H2O_data_parsing_timezone:	UTC
H2O_cluster_version:	3.36.1.1
H2O_cluster_version_age:	7 days, 19 hours and 24 minutes
H2O_cluster_name:	H2O_from_python_Gilseung_ax8c8b
H2O_cluster_total_nodes:	1
H2O_cluster_free_memory:	1.914 Gb
H2O_cluster_total_cores:	8
H2O_cluster_allowed_cores:	8
H2O_cluster_status:	locked, healthy
H2O_connection_url:	http://localhost:54321
H2O_connection_proxy:	{"http": null, "https": null}
H2O_internal_security:	False
Python_version:	3.8.8 final

- init 메서드를 사용하면 h2o에서 사용할 클러스터의 가동 시간, 시간대, 할당된 컴퓨팅 자원과 관련된 정보가 함께 출력됨
- H2O AutoML의 환경과 현재 사용자 수 등에 따라 해당 정보가 바뀔 수 있습니다.

예제 데이터 불러오기

h2o를 사용해 학습할 예제 회귀 데이터를 불러옵니다.

예제 데이터 불러오기

```
1  import pandas as pd
2  from sklearn.model_selection import train_test_split
3  df = pd.read_csv("../../data/regression/wankara.csv")
4  train_df, test_df = train_test_split(df)
```

- **라인 4:** 특징과 라벨을 분리하지 않은 채로 학습 데이터와 평가 데이터로 분할했습니다. 그 이유는 H2OAutoML에서는 H2OFrame이라는 데이터프레임에 특징과 라벨이 모두 포함돼야 모델을 학습할 수 있기 때문입니다.

H2OAutoML 클래스

h2o에서는 H2OAutoML이라는 클래스를 사용해 머신러닝 자동화를 수행하는 인스턴스를 만들 수 있습니다.

주요 인자

인자	설명
max_runtime_secs	최대 학습 시간(초)
max_models	고려하는 모델 수
max_runtime_secs_per_model	모델별 최대 학습 시간(초)
exclude_algos	학습하지 않을 모델 목록

인스턴스 생성

```
1 from h2o.automl import H2OAutoML
2 aml = H2OAutoML(max_runtime_secs = 3600)
```

train 메서드

H2OAutoML의 인스턴스는 train이라는 메서드를 사용해 여러 모델을 학습하고 평가합니다.

train 메서드의 주요 인자

인자	설명
x	특징 이름 목록
y	라벨 이름
training_frame	학습 데이터 (H2O Frame)
leaderboard_frame	검증 데이터 (H2O Frame)

H2O Frame으로 변환

```
1  train_df = h2o.H2OFrame(train_df)
2  test_df = h2o.H2OFrame(test_df)
```

```
Parse progress: |████████████████████████████████████████████████████████████████| (done) 100%
Parse progress: |████████████████████████████████████████████████████████████████| (done) 100%
```

train 메서드 (계속)

H2OAutoML의 인스턴스는 train이라는 메서드를 사용해 여러 모델을 학습하고 평가합니다.

회귀모델 학습 자동화

```
1 aml.train(x = list(df.drop('y', axis = 1).columns),
2         y = 'y',
3         training_frame = train_df,
4         leaderboard_frame = test_df)
```

```
Model Details
==============
H2OStackedEnsembleEstimator :  Stacked Ensemble
Model Key:  StackedEnsemble_BestOfFamily_6_AutoML_1_20220421_113737

No model summary for this model

ModelMetricsRegressionGLM: stackedensemble
** Reported on train data. **

MSE: 0.7639552304417417
RMSE: 0.8740453251644
... 중략 ...

ModelMetricsRegressionGLM: stackedensemble
** Reported on cross-validation data. **

MSE: 1.395131560910733
RMSE: 1.1811568739632907... 중략 ...
```

		mean	sd	cv_1_valid	cv_2_valid	cv_3_valid	cv_4_valid	cv_5_valid
0	mae	0.907263	0.053535	0.961983	0.881519	0.967105	0.850988	0.874717
1	mean_residual_deviance	1.388473	0.131406	1.501449	1.187794	1.512434	1.368019	1.372668
2	mse	1.388473	0.131406	1.501449	1.187794	1.512434	1.368019	1.372668
3	null_deviance	57349.900000	4023.832800	64409.450000	55869.168000	56147.594000	54236.273000	56087.008000
4	r2	0.994116	0.000593	0.994055	0.995009	0.993344	0.994149	0.994024
5	residual_deviance	335.732700	43.886790	382.869500	276.756070	373.571260	311.908400	333.558200
6	rmse	1.177248	0.056573	1.225336	1.089860	1.229811	1.169623	1.171609
7	rmsle	0.028345	0.003092	0.032939	0.026443	0.030140	0.026432	0.025770

모델 평가

H2O Frame인 리더보드는 leaderboard 속성을 이용해 확인할 수 있으며 head 메서드를 사용해 일부를 확인할 수 있습니다.

회귀모델 학습 결과 확인

```
1 lb = aml.leaderboard
2 display(lb.head(rows = 10))
```

model_id	rmse	mse	mae	rmsle	mean_residual_deviance
StackedEnsemble_BestOfFamily_6_AutoML_1_20220421_113737	1.2228	1.49524	0.944005	0.0305209	1.49524
StackedEnsemble_BestOfFamily_4_AutoML_1_20220421_113737	1.24926	1.56066	0.957628	0.0310331	1.56066
StackedEnsemble_AllModels_6_AutoML_1_20220421_113737	1.25538	1.57598	0.979048	0.0293236	1.57598
DeepLearning_grid_1_AutoML_1_20220421_113737_model_1	1.27578	1.62762	0.978258	0.0321887	1.62762
StackedEnsemble_BestOfFamily_5_AutoML_1_20220421_113737	1.35422	1.8339	1.04192	0.0392248	1.8339
DeepLearning_grid_1_AutoML_1_20220421_113737_model_4	1.37484	1.89018	1.06594	0.0340014	1.89018
DeepLearning_grid_1_AutoML_1_20220421_113737_model_30	1.37948	1.90296	1.04116	0.0300074	1.90296
StackedEnsemble_BestOfFamily_2_AutoML_1_20220421_113737	1.40048	1.96133	1.06444	0.0356848	1.96133
StackedEnsemble_BestOfFamily_3_AutoML_1_20220421_113737	1.40211	1.96592	1.09454	0.0347414	1.96592
StackedEnsemble_AllModels_2_AutoML_1_20220421_113737	1.40431	1.97208	1.09261	0.0347867	1.97208

- BestofFamily_6 스태킹 앙상블의 성능이 가장 좋았음
- 상위 10개 모델이 모두 스태킹 앙상블 혹은 딥러닝 모델임

모델 활용

리더보드에서 가장 점수가 높은 모델을 이용해 라벨을 예측하려면 predict 메서드를 사용합니다. 단, 입력하는 데이터 역시 H2OFrame이어야 합니다.

회귀모델 활용

```
1 aml.predict(test_df)
```

predict
35.4338
36.0951
72.4443
30.7745
70.8613
49.8148
56.9813
66.2821
76.4649
51.9912