Online 기반 Random forest 알고리즘

20510082 전우진

20510083 한재웅

CONTENTS

1. 데이터

Rotating Hyperplane dataset 사용

- Gradual concept drift가 일어나는 parameter를 조절해서 만든 가상의 데이터셋

DataSet	차원	observation	output class	proportion (1:0)	K
1	10	10,000	2	5027:4974	2
2	10	10,000	2	5017:4983	5

K: corresponding weight가 drift 될 차원의 수

attr0	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	output
0.031337	0.566490	0.389565	0.424643	0.853130	0.552442	0.651868	0.258164	0.688760	0.107980	0
0.736709	0.976751	0.666018	0.330315	0.348992	0.613775	0.391591	0.028845	0.892445	0.944336	1
0.296464	0.774811	0.681787	0.736286	0.571683	0.931030	0.614503	0.485105	0.686068	0.191413	1
0.268038	0.232699	0.930201	0.947699	0.412382	0.653455	0.348090	0.999704	0.025922	0.882774	1
0.773298	0.265773	0.995941	0.149586	0.017813	0.429112	0.201519	0.308675	0.655395	0.300896	0
...	...	...	...	...	...	...	...	...	...	...
0.862055	0.338475	0.970367	0.089220	0.134278	0.788218	0.437395	0.328031	0.657321	0.422218	1
0.066373	0.756530	0.936229	0.227335	0.613399	0.691645	0.265779	0.062472	0.490900	0.806306	0
0.712069	0.269874	0.687450	0.413940	0.034405	0.019472	0.281914	0.306431	0.308301	0.049732	0
0.794370	0.246686	0.239883	0.208131	0.816414	0.764663	0.135609	0.381933	0.625414	0.030935	1
0.338349	0.427789	0.121447	0.828973	0.023751	0.457726	0.415431	0.275512	0.544775	0.817484	0

rows × 11 columns

Hyperplane 3

	attr0	attr1	attr2	attr3	attr4	attr5	attr6	attr7	attr8	attr9	output
0	0.588680	0.994291	0.692638	0.492426	0.244472	0.737020	0.681701	0.181916	0.655073	0.162986	1
1	0.761392	0.987276	0.767314	0.125468	0.705822	0.962866	0.418334	0.347568	0.443218	0.939095	1
2	0.200499	0.346797	0.198434	0.832916	0.743052	0.334474	0.796501	0.002093	0.295237	0.445850	0
3	0.726625	0.936809	0.103744	0.516967	0.683031	0.333098	0.635495	0.848070	0.263577	0.174056	0
4	0.587769	0.010560	0.716553	0.826589	0.896887	0.118905	0.405279	0.530525	0.430691	0.991987	1
...	...	...	...	...	...	...	...	...	...	...	...
9995	0.400476	0.366739	0.014778	0.392485	0.512218	0.202046	0.437779	0.768193	0.615782	0.721150	1
9996	0.249982	0.820053	0.564774	0.271589	0.657724	0.731979	0.811500	0.463711	0.914278	0.713895	1
9997	0.409096	0.228425	0.607483	0.325246	0.756762	0.052312	0.497422	0.154382	0.170557	0.614689	1
9998	0.060611	0.570429	0.590032	0.569474	0.304741	0.760363	0.253692	0.347916	0.809181	0.304610	1
9999	0.661866	0.413878	0.764754	0.308570	0.732954	0.866010	0.442453	0.882727	0.309690	0.342325	0

10000 rows × 11 columns

Hyperplane 6

2. 실험

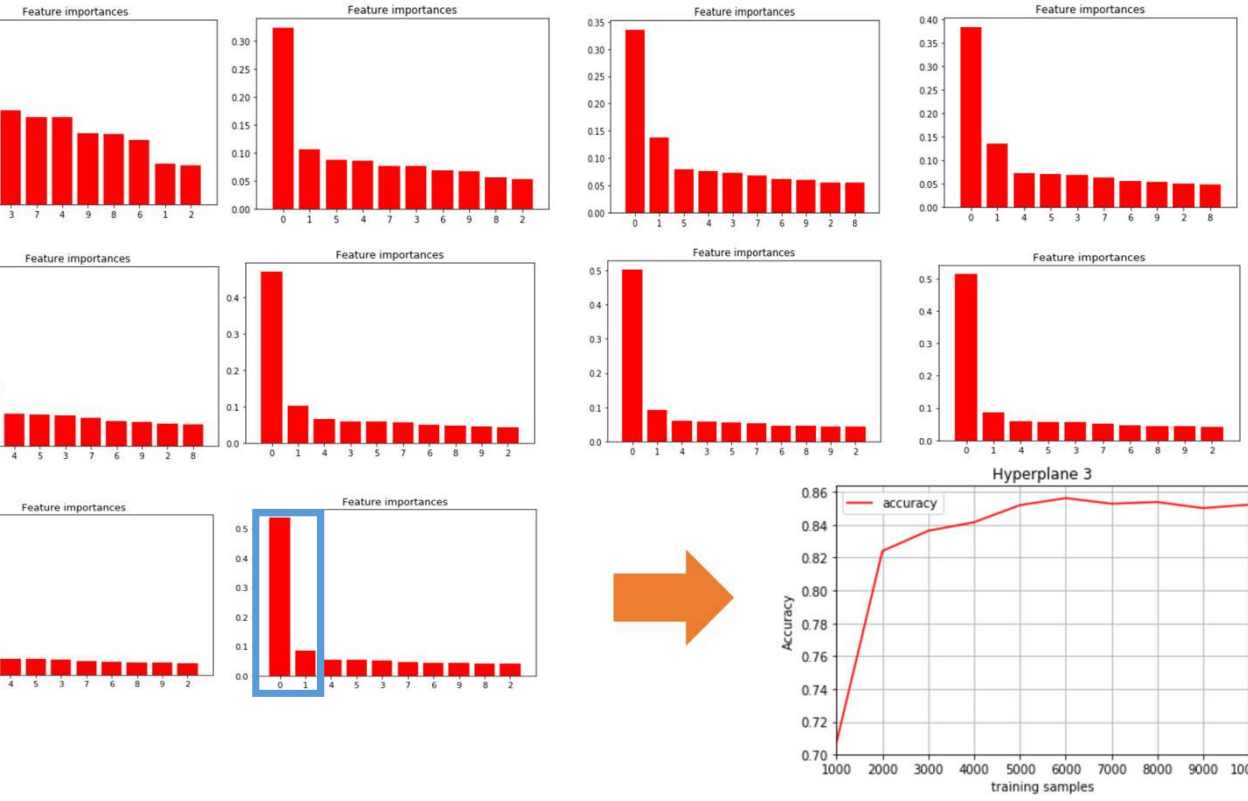
설계

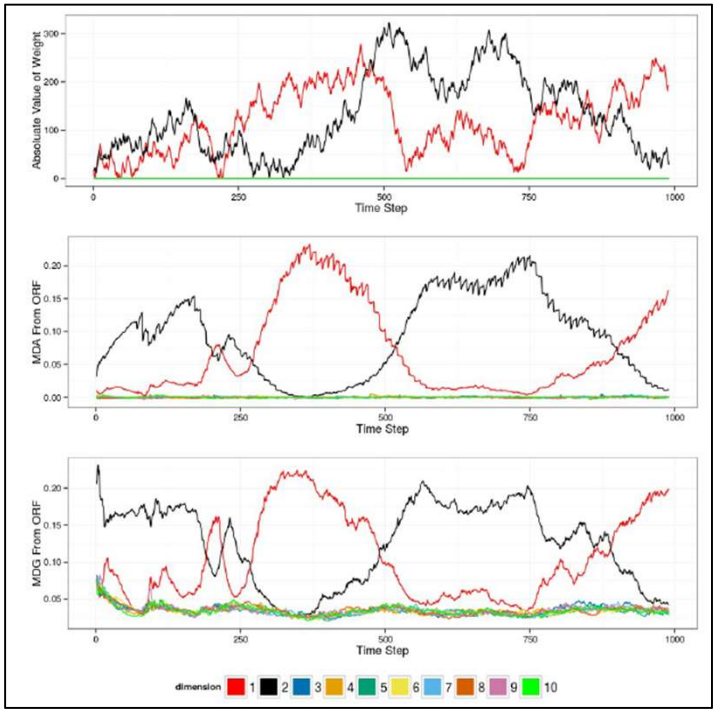
- Online random forest algorithm
  - ➢ 데이터셋을 training과 testing dataset으로 분리
  - ➢ Incremental Tree Building을 통해 adaptive learning이 될 수 있도록 적용시킴
    - ✓ RF Classifier에는 online learning method를 지원하지 않음
    - ✓ Parmeter 중 하나인 warm_start를 True로 둠으로써 최대한 비슷한 효과를 가져오도록 함
  - ➢ Online Bootstrap Aggregation을 통해 무작위 표본을 추출하고 training dataset을 훈련시키고 testing dataset을 통해 정확도 및 오류율을 추정함
  - ➢ Tree discarding을 통해 정확도가 높게 나타나는 구간을 확인하는 앙상블 구조조정의 형태를 최종적으로 거치게 됨

2. 실험

결과

- Hyperplane 3 (N: 10000, 1000개씩 분할 및 n_estimator +10)
  - 정확도 84.58%
  - Gradual concept drift가 일어나는 변수에 대한 Feature importance가 높다는 것을 확인함

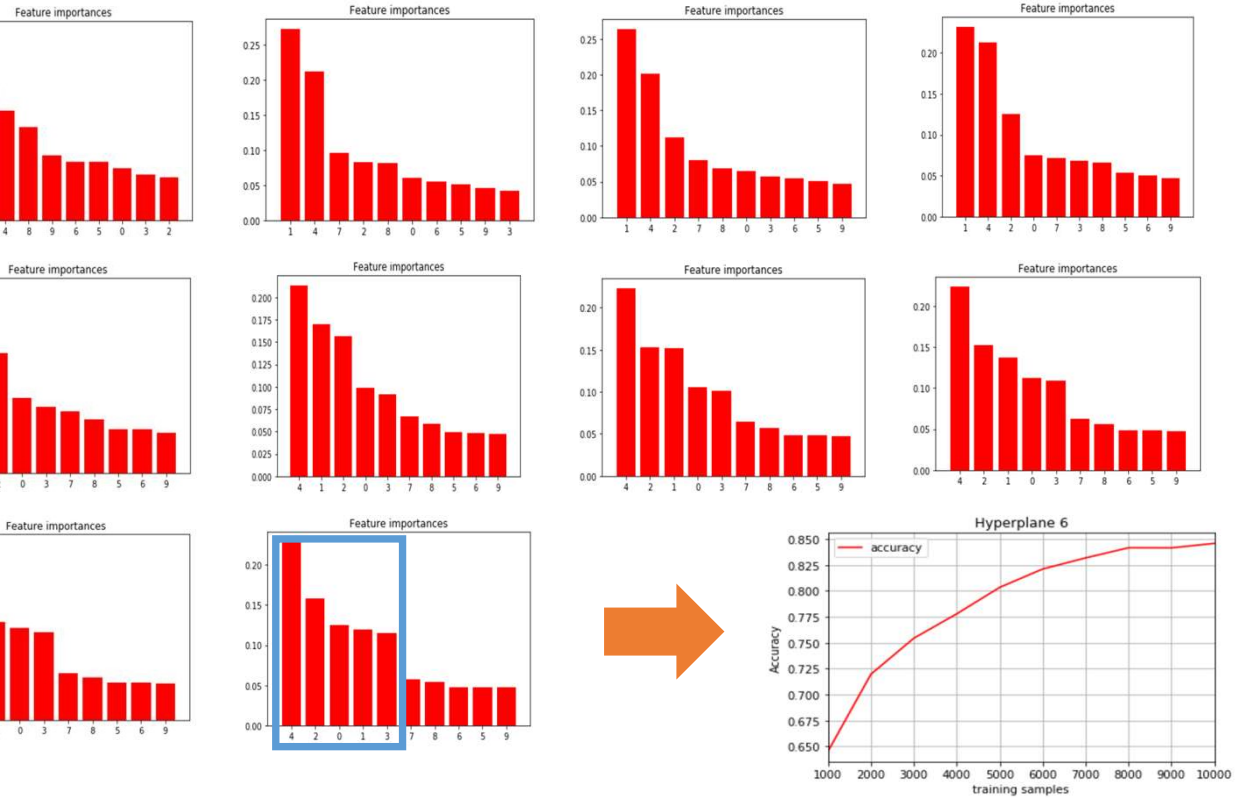


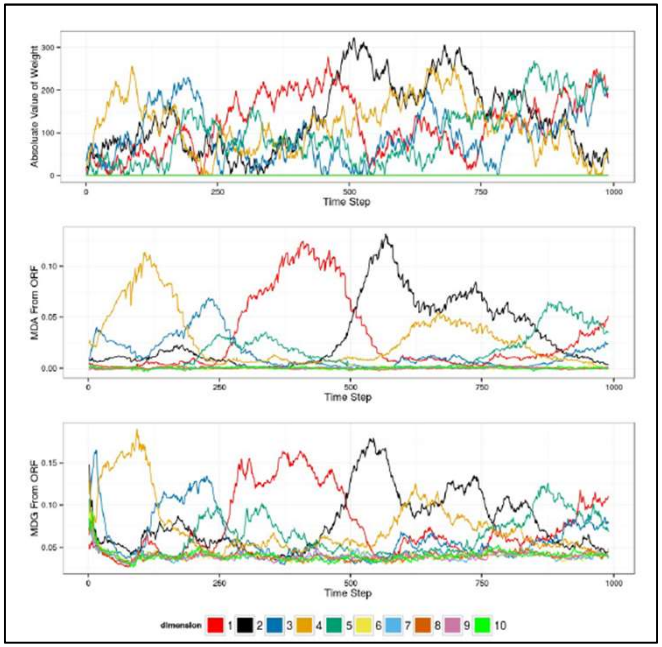


2. 실험

결과

- Hyperplane 6 (N: 10000, 1000개씩 분할 및 n_estimator +10)
  - 정확도 85.22%
  - Gradual concept drift가 일어나는 변수에 대한 Feature importance가 높다는 것을 확인함





3. 평가

결과

- Concept Drift data를 사용한 해당 논문에서의 결과에서 나타난 gradual concept drift를 확인함
- 해당 논문에서 concept drift가 일어난 변수들의 움직임과 feature importance에서 중요한 변수로 나타난 것과 일치함
  - Hyperplane 3: 변수 1, 2
  - Hyperplane 6: 변수 1, 2, 3, 4, 5
- 1000개씩 training을 시키면서 n_estimator의 개수를 조금씩 증가시키면서 정확도가 개선되는 모습을 확인함

평가 방법

- Hyperplane 3 Accuracy: 84.58%
- Hyperplane 6 Accuracy: 85.22%

한계점

- 기존에 계획했던 online 학습방식의 random forest의 구현을 하지 못함
- 논문에서 online random forest의 특징 중 하나로 주장했던 tree discarding을 적용하지 못함
- Online random forest 적합한 데이터를 찾기 어려워서 다양한 데이터에 적용시키지 못함

Thank you