A Feature Selection Study for Medical Image deep learning

Sanghyuck Lee

2021. 06 15.

Outline

- Deep Feature Selection study
- 중견 연구 제안서
- 갑상선 연구
- 산학 협력 프로젝트

Experiments Setting: Dataset

~~1. UJIndoorLoc~~

~~2. slice_localization_data~~

~~3. blogData~~

4. CommViolPredUnnormalizedData

5. communities

6. Residential-Building-Data-Set

~~7. c2k_data_comma~~

8. YearPredictionMSD

9. CNNpred

10. superconduct

11. BuzzInSocialMedia_TomsHardware

12. Geographical Original of Music

13. OnlineNewsPopularity

~~14. Facebook Comment Volume Dataset~~

15. wiki4HE

~~16. WECs~~

17. cancer mortality rates for US counties

18. Insurance Company Benchmark (COIL 2000)

19. ailerons

20. tic

21. puma32h

~~22. PoleTelecomm~~

23. Bank

24. Reaction Network (Undirected)

25. waveform-5000

26. Toy 1 (n=500, p=200)

27. Toy 2 (n=1000, p=400)

28. Toy 3 (n=1500, p=600)

Experiments Setting: Dataset

Dataset Name	(#instances, # features)
4. CommViolPredUnnormalizedData	(2214, 145)
5. communities	(1993, 127)
6. Residential-Building-Data-Set	(371, 105)
8. YearPredictionMSD	(515344, 91)->(20000, 91)
9. CNNpred	(9919, 84)
10. superconduct	(21262, 82) ->(20000, 82)
11. BuzzInSocialMedia_TomsHardware	(28178, 97) -> (20000, 97)
12. Geographical Original of Music	(1058, 70)
13. OnlineNewsPopularity	(39643, 60) -> (20000, 60)
15. wiki4HE	(912, 46)
17. cancer mortality rates for US counties	(3046, 32)

Experiments Setting: Dataset

Dataset Name	(#instances, # features)
18. Insurance Company Benchmark (COIL 2000)	(5821, 86)
19. ailerons	(13749, 41)
20. tic	(9821, 86)
21. puma32h	(8191, 33)
23. Bank	(4499, 33)
24. Reaction Network (Undirected)	(65553, 28) -> (20000, 28)
25. waveform-5000	(4999, 41)
26. Toy1	(500,200)
27. Toy2	(1000,400)
28. Toy2	(1500,600)

Experiments Setting: Methods

- Experiments methods: use Ordinary Least Squares (OLS) for regressor
    - Knockoff Filter (Fixed-X):
        - Equicorrelated method for S matrix
        - Lasso coefficients for Feature importances
        - Signed maximum for Feature statistics
        - FDR level: 0.25
    - No Filtering (use all predictors for regressor)
    - Correlation Coefficients Sequential Forward Selection
        - Calculate Correlation Coefficients between each x and y
        - Predict p times (e.g. if p=10, 1(best)->2->...->9->10(worst))
        - Use best case performance as final performance

Experiments Results

Method	FX-Knockoff		Not filtering		CC-SFS	
Dataset	CC	RMSE	CC	RMSE	CC	RMSE
4. CommViolPred (2214, 145)	**0.974**	443.462	0.968	488.931	**0.974**	**438.117**
5. Communities (1993, 127)	**0.654**	**0.139**	-4.890	0.351	0.651	0.140
6. Residential-Building (371, 105)	**0.885**	**53.957**	0.916	47.297	**0.972**	**28.340**
8. YearPredictionMSD (20000, 91)	**0.333**	**18.422**	0.334	18.417	**0.337**	**18.378**
9. CNNpred (9919, 84)	**0.366**	**0.005**	0.842	0.002	**0.842**	**0.002**
10. Superconduct (20000, 82)	**0.000**		0.740	17.404	**0.740**	**17.403**
11. BuzzInSocialMedia (20000, 97)	0.802	**4835.44**	0.956	2335.21	**0.959**	**2267.64**
12. Geographical Music (1058, 70)	0.189	45.146	0.186	45.238	**0.222**	**44.278**
13. OnlineNewsPopularity (20000, 60)	0.019	8807.82	0.013	8835.52	**0.022**	**8797.20**
15. wiki4HE (912, 46)	**0.221**	**1.182**	0.228	1.177	**0.239**	**1.168**
17. cancer mortality (3046, 32)	**0.137**	1.822	0.138	1.822	**0.142**	**1.817**

Experiments Results

Method	FX-Knockoff		Not filtering		CC-SFS	
Dataset	CC	RMSE	CC	RMSE	CC	RMSE
18. Insurance (5821, 86)	0.037	0.231	0.025	0.232	**0.044**	**0.230**
19. Ailerons (13749, 41)	**0.504**	0.000	0.817	0.000	**0.817**	**0.000**
20. Tic (9821, 86)	**0.035**	0.236	0.039	0.236	**0.040**	**0.235**
21. puma32h (8191, 33)	0.212	**0.027**	0.207	**0.027**	**0.213**	**0.027**
23. Bank (4499, 33)	0.513	**0.084**	0.513	**0.084**	**0.516**	**0.084**
24. Reaction Network (20000, 28)	0.983	**2.337**	0.983	2.311	**0.985**	**2.198**
25. waveform-5000 (4999, 41)	**0.458**	**0.599**	0.455	0.600	0.457	0.600
toy1(500,200)	0.934	1.220	0.912	1.421	**0.950**	**1.085**
toy2(1000,400)	0.966	1.231	0.951	1.482	**0.968**	**1.192**
toy2(1500,600)	**0.975**	**1.169**	0.961	1.477	**0.975**	1.172

Discussion

- Not working on some datasets

  => Needs a way to satisfy mathematical conditions

  (e. g. Sigma inverse, S matrix)

- A Study on the Main Advantages of Knock-Off Filters

- High dimensional case issues

  => Fixed-X -> n>2p

  => Model-X -> any cases

중견 연구 제안서 작성

- 가제:

  한글: 의료 분야내 주요 관심 구조물 세그멘테이션을 위한 사전훈련 최신 딥러닝 모델 개발

  영문: Development of Pre-trained Deep Learning Model for Segmentation of Structures of Major Interest in the Medical Field

- 연구 목표:

  본 과제에서는 몇몇 주요한 해부학적 구조에 대해서 세계 최고 수준의 성능을 갖는 사전학습된 세그멘테이션 모델을 배포함으로써, 다양한 의료분야 응용에 성공적으로 기여하는 것을 최종 목표로 하며, 이를 위해 다음과 같은 연차별 연구 목표를 설정하여 추진한다.

진행하고자 하는 계획

- 연차별 계획 수립을 위한 적절한 주요 관심 구조물 선정:

  - 눈, 담낭 / 심장, 치아
  - 복부, 뇌, 유방, 심장, 가슴, 좌심실, 간, 폐, 전립선, 피부, 복합

  등 다양한 public 데이터셋이 있는 것으로 확인,

  해당 데이터셋의 형태를 확인 및 확보할 계획

  => 실제로 확보 가능한 데이터셋 리스트 작성

  => 리스트 내에서 적절한 관심 구조물 선정

  => 각 구조물 당, 최신 기술 조사 후 연차 계획서 작성

갑상선 연구

- 완료 현황: 각각 1323명 환자 (기존에는 1267 환자)

유형	눈 (L)	눈 (R)	시신경 (L)	시신경 (R)	내직근 (L)	내직근 (R)	외직근 (L)	외직근 (R)	상직근 (L)	상직근 (R)	하직근 (L)	하직근 (R)	상안검 (L)	상안검 (R)	지방 (L)	지방 (R)	눈물주머니 (L)	눈물주머니 (R)	안와 (L)	안와 (R)	평균
AX 1	0.85	0.81	0.42	0.47	0.64	0.63	0.73	0.66							0	0					0.52
AX 2	0.91	0.89	0.75	0.72	0.93	0.93	0.93	0.9							0	0					0.70
AX 3	0.96	0.96	0.95	0.94	0.95	0.96	0.94	0.95							0.08	0.08					0.78
AX 4	0.94	0.94	0.82	0.84	0.82	0.84	0.79	0.75							0	0					0.67
AX 5	0.81	0.81	0.36	0.37	0.34	0.36	0.32	0.38							0	0					0.38
AX 6																	0	0			0.00
CO 1	0.96	0.96																			0.96
CO 2			0.89	0.89	0.91	0.91	0.91	0.89	0.86	0.9	0.91	0.88									0.90
CO 3																			0.08	0.08	0.08
SA 1		0.96		0.92						0.96		0.93		0.96							0.95
SA 2	0.96		0.92						0.96		0.93		0.96								0.95
계																					0.67

0% 100%

산학 협력 프로젝트

- 제출 실적
  - SW 제작 (산학 1), 국제학술대회 논문 1편 이상 게재 (산학 2), 국내 특허출원 1건 이상 (산학 2)
- 현재 진행 사항
  - SW 등록: 진행중
  - 특허 출원: 기존 논문 번역본 제작 중
  - 투고 저널 리스트업 완료

References

Barber, Rina Foygel, and Emmanuel J. Candès. "Controlling the false discovery rate via knockoffs." *Annals of Statistics* 43.5 (2015): 2055-2085.

Benjamini, Yoav, and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995): 289-300.

Chen, Jiajie, Anthony Hou, and Thomas Y. Hou. "Some analysis of the knockoff filter and its variants." *arXiv preprint arXiv:1706.03400* (2017).

Jordon, James, Jinsung Yoon, and Mihaela van der Schaar. "KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks." *International Conference on Learning Representations*. 2018.

Sarkar, Sanat K., and Cheng Yong Tang. "Adjusting the Benjamini-Hochberg method for controlling the false discovery rate in knockoff assisted variable selection." *arXiv preprint arXiv:2102.09080* (2021).