Active feature acquisition via explainability-driven ranking

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, Vijaya B. Kolachalama

ICML
International Conference
On Machine Learning

Problem statement

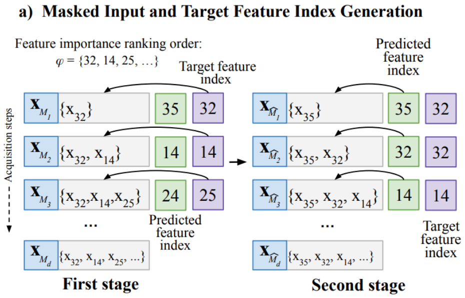
- Real-world feature acquisition is often costly, time-consuming, and sequential. Active feature acquisition (AFA) frameworks address this sequential optimization problem.
- The objective is to find a predictor f_θ and a policy network q_π such that the given constraint objective is minimized:

$$\min_{\theta,\pi} \mathbb{E}_{\mathbf{x}yk}\mathbb{E}_{M\sim q_\pi}[\ell(f_\theta(\mathbf{x}_M), y)], \text{ s.t.} \sum_{j\in M} c_j \leq k.$$

- Traditionally, this problem has been addressed using RL-based algorithms or greedy methods based on information theory.
- We developed a method that leverages local explanation techniques to generate instance-specific feature importance rankings, by reframing the AFA problem as a feature prediction task.

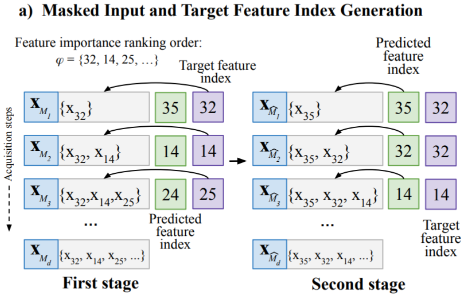
Our method

- We use a two-step training strategy.
- First, we trained a classifier and employed a feature explanation method to derive importance rankings.
- In the first stage, we fed the masked input using features ordered by their importance rankings, where the target is the next feature in the ranking sequence.



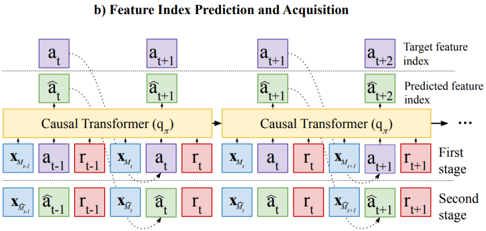
Our method

- During inference, q_π, is not % 100 accurate, so the feature subset $\hat{M}_t$, generated by q_π, does not always contain the top t features with the highest ranking order.
- To address this, in the second stage, we generated a mask from the policy predictions and selected the target feature as the highest-ranked unacquired feature.

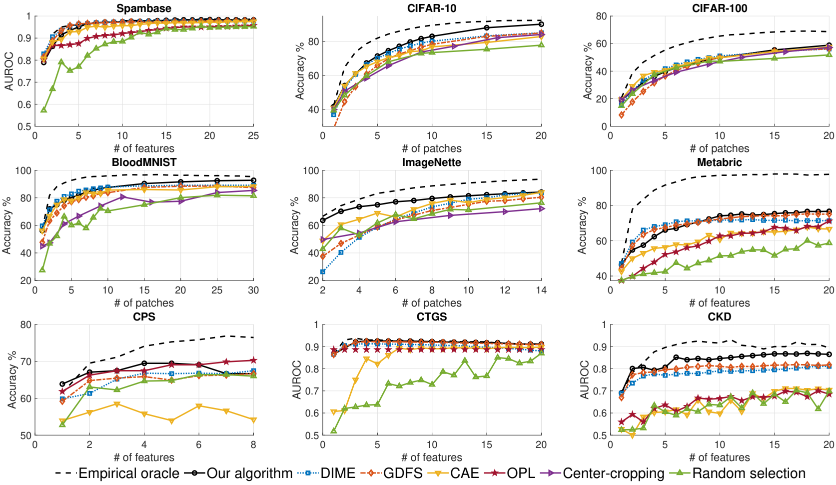


Our method

- A decision transformer was employed [1] as the policy network.
- At each time step, it receives three tokens: the masked input, an action token, and a reward token. The action token represents the index of the most recently acquired feature, while the reward corresponds to the predictor network's output.



Results



Spambase
AUROC
# of features
CIFAR-10
Accuracy %
# of patches
CIFAR-100
Accuracy %
# of patches
BloodMNIST
Accuracy %
# of patches
ImageNette
Accuracy %
# of patches
Metabric
Accuracy %
# of features
CPS
Accuracy %
# of features
CTGS
AUROC
# of features
CKD
AUROC
# of features
Empirical oracle
Our algorithm
DIME
GDFS
CAE
OPL
Center-cropping
Random selection

Results

Table: Stage-wise classification results, with extended first-stage training (250 epochs), demonstrate the advantage of our two-stage approach over prolonged single-stage training.

	CIFAR10	**CIFAR100**	**BloodMNIST**	**ImageNette**
# of classes:	10	100	8	10
First-stage (250)	$75.96_{\pm 0.16}\%$	$45.91_{\pm 0.36}\%$	$79.83_{\pm 0.19}\%$	$73.95_{\pm 0.25}\%$
First-stage	$75.76_{\pm 0.19}\%$	$46.05_{\pm 0.25}\%$	$79.25_{\pm 0.15}\%$	$73.76_{\pm 0.42}\%$
Second-stage	$78.44_{\pm 0.15}\%$	$46.99_{\pm 0.15}\%$	$83.87_{\pm 1.05}\%$	$78.96_{\pm 0.12}\%$

	Spam	**Metabric**	**CPS**	**CTGS**	**CKD**
# of classes:	2	6	3	2	2
First-stage (250)	$0.952_{\pm .001}$	$62.52_{\pm 1.27}\%$	$67.23_{\pm 0.48}\%$	$0.916_{\pm .0002}$	$0.822_{\pm .01}$
First-stage	$0.951_{\pm .0002}$	$62.48_{\pm 1.39}\%$	$67.21_{\pm 0.15}\%$	$0.916_{\pm .0004}$	$0.825_{\pm .008}$
Second-stage	$0.955_{\pm .0001}$	$69.83_{\pm 0.41}\%$	$67.45_{\pm 0.13}\%$	$0.916_{\pm .0001}$	$0.836_{\pm .07}$

Conclusions

- Our method outperforms or matches state-of-the-art AFA approaches.
- Instance-specific feature importance rankings derived from local explanation methods are effective for the AFA task.
- Two-stage training strategy is effective.

Acknowledgments

- This project was supported by grants from the National Institute on Aging's Artificial Intelligence and Technology Collaboratories (P30-AG073105), the American Heart Association (20SFRN35460031), and the National Institutes of Health (R01-HL159620, R01-AG062109, R01-AG083735, and R01-NS142076).

References

[1] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 15084–15097, Curran Associates, Inc., 2021.