Online Semi-Supervised Learning in Contextual Bandits with Episodic Reward

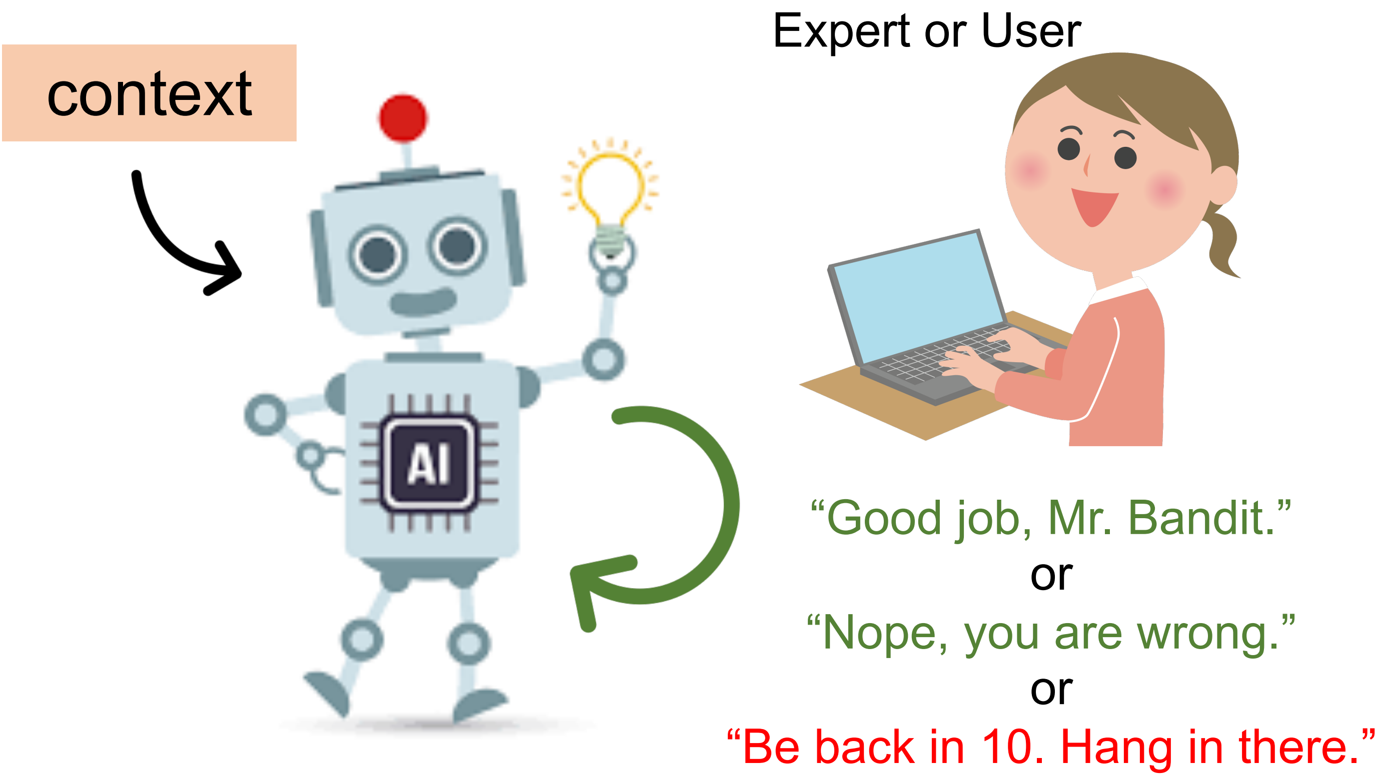
Baihan Lin

Columbia University

baihan.lin@columbia.edu

Online Learning in Real World

Rewards usually come in episodes…



In scenarios like

- Inactivity in Rec Sys
- Personalized medicine
- Intelligent systems
- Ads in on/off seasons

Contextual Bandits with Episodic Reward

Algorithm 1 Online Learning with Episodic Reward

for t = 1,2,3,$\cdots$, T **do**
  $(\mathbf{x}(t), \mathbf{r}(t))$ is drawn according to $\mathbb{P}_{x,r}$
  Context $\mathbf{x}(t)$ is revealed to the player
  Player chooses an action $a_t = \pi_t(\mathbf{x}(t))$
  Feedback $r_{a_t,t}(t)$ for only chosen arms are episodically revealed
  Player updates its policy π_t
end for

given by a probability $p_r \in [0, 1]$

Semi-Supervised Solution

We proposed Background Episodically Rewarded LinUCB (*BerlinUCB*)

In episodes where there is no feedbacks, we use two strategies

- Only update covariance matrices
- Create pseudo-rewards with self-supervision

Self-supervision modules via clustering

- Gaussian mixture model (GMM)
- K-means
- K nearest neighbors (KNN)

Algorithm 2 BerlinUCB

Initialize $c_t \in \mathbb{R}_+, \mathbf{A}_a \leftarrow \mathbf{I}_d, \mathbf{b}_a \leftarrow \mathbf{0}_{d\times 1} \forall a \in \mathcal{A}_t$
for t = 1,2,3,$\cdots$, T **do**
  Observe features $\mathbf{x}_t \in \mathbb{R}^d$
  for all $a \in \mathcal{A}_t$ **do**
    $\hat{\theta}_a \leftarrow \mathbf{A}_a^{-1}\mathbf{b}_a$
    $p_{t,a} \leftarrow \hat{\theta}_a^\top \mathbf{x}_t + c_t\sqrt{\mathbf{x}_t^\top \mathbf{A}_a^{-1}\mathbf{x}_t}$
  end for
  Choose arm $a_t = \arg\max_{a\in\mathcal{A}_t} p_{t,a}$
  if the background revealed the feedbacks **then**
    Observe feedback $r_{a_t,t}$
    $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_t\mathbf{x}_t^\top$
    $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_{a_t,t}\mathbf{x}_t$
  elif the background revealed NO feedbacks **then**
    if use self-supervision feedback
      $r' = [a_t == \text{predict}(\mathbf{x}_t)]$
      $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r'\mathbf{x}_t$
    elif
      $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_t\mathbf{x}_t^\top$
    end if
  end if
end for

Empirical Evaluations

Evaluate performance in online classification task with bandit feedback

Datasets

- MNIST
- Warfarin

Nonstationary Environments

- Nonstationary Context: varying cluster distribution
- Nonstationary Context: negative images
- Nonstationary Reward: shuffled class labels
- Nonstationary Reward: Multi-Task Environment
- Nonstationary Oracle: Fixed vs. Extendable Arms
- Nonstationary Oracle: Varying Episodic Rewards

Results

Table 1: **Accuracy:** Stationary contexts with different probabilities of reward revealing

	MNIST (varying p_r)		MNIST (p_r=0.5)		MNIST (p_r=0.1)		MNIST (p_r=0.01)	
	fixed arms	extendable	fixed arms	extendable	fixed arms	extendable	fixed arms	extendable
LinUCB	0.1134	0.094	0.1080	0.0962	0.0842	0.0762	0.0252	0.0902
BerlinUCB	0.1138	0.0990	0.1102	0.0990	0.1016	0.0926	0.0788	0.0896
B-Kmeans	0.2594	0.2130	0.2674	**0.2678**	**0.3132**	**0.2760**	**0.1400**	0.0828
B-KNN	**0.2642**	**0.2398**	**0.2722**	0.2642	0.2954	0.2622	0.1384	**0.0938**
B-GMM	0.0958	0.0768	0.1060	0.0728	0.0958	0.0320	0.1034	0.0120

Table 2: **Accuracy:** Nonstationary cases with different probabilities of reward revealing

	MNIST - varying clusters			MNIST - negative images			MNIST - shuffled rewards		
Fixed Arms	p_r=0.5	p_r=0.1	p_r=0.01	p_r=0.5	p_r=0.1	p_r=0.01	p_r=0.5	p_r=0.1	p_r=0.01
LinUCB	**0.1086**	0.0984	0.0690	0.1044	0.0920	0.0732	0.1172	0.0832	0.0512
BerlinUCB	0.1024	0.1002	0.0828	0.1016	0.0970	0.1016	0.1120	0.0966	0.0832
B-Kmeans	0.0952	**0.1096**	**0.1074**	0.1082	0.1036	0.0768	**0.2776**	0.2878	**0.1618**
B-KNN	0.0984	0.1002	0.1050	**0.1762**	**0.1732**	0.1016	0.2600	**0.3162**	0.1556
B-GMM	0.1072	0.0974	0.1034	0.0954	0.1074	**0.1018**	0.1158	0.1172	0.1062

	MNIST - varying clusters			MNIST - negative images			MNIST - shuffled rewards		
Extendable Arms	p_r=0.5	p_r=0.1	p_r=0.01	p_r=0.5	p_r=0.1	p_r=0.01	p_r=0.5	p_r=0.1	p_r=0.01
LinUCB	0.0994	0.0930	0.0866	0.0888	0.0918	**0.1014**	0.1010	0.0872	0.1014
BerlinUCB	0.0930	0.0946	**0.0938**	0.0924	0.0930	0.0910	0.0994	0.0926	0.0918
B-Kmeans	**0.1010**	**0.0990**	0.0380	0.0910	0.0712	0.0308	0.2478	0.2336	0.0654
B-KNN	0.0958	0.0970	0.0570	**0.1780**	**0.1228**	0.0460	**0.2606**	**0.2818**	**0.1016**
B-GMM	0.0756	0.0296	0.0108	0.0680	0.0338	0.0118	0.0720	0.0258	0.0180

Results

Table 3: **Accuracy:** Nonstationary contexts with varying episodic rewards

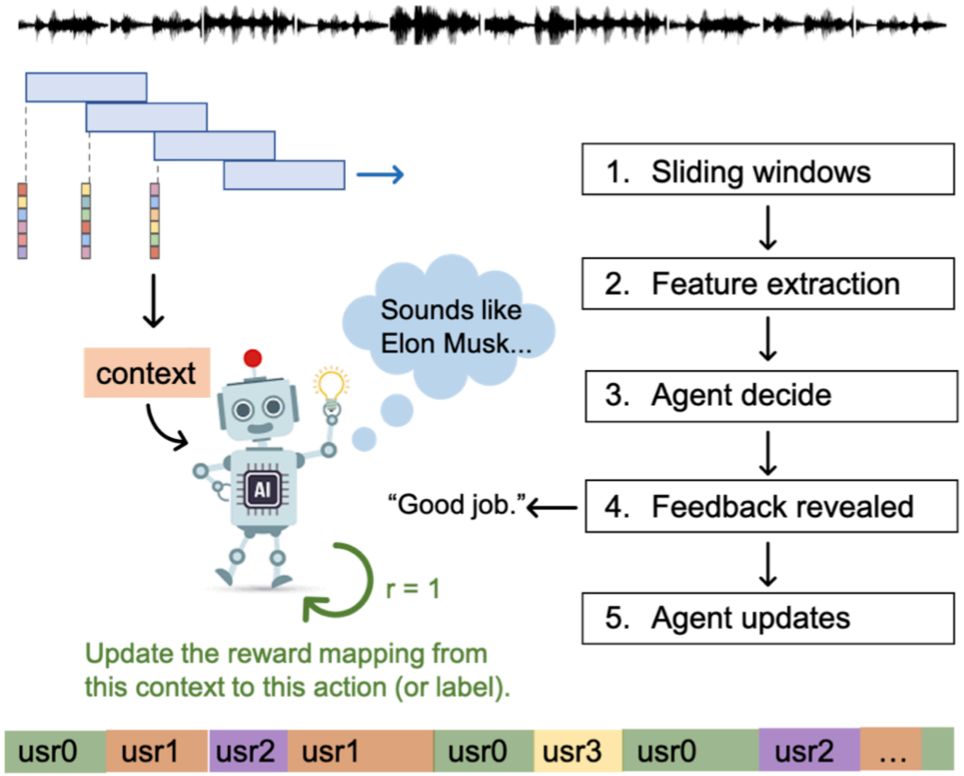
	varying cluster distribution				negative images				average
	MNIST-F	MNIST-E	Warfarin-F	Warfarin-E	MNIST-F	MNIST-E	Warfarin-F	Warfarin-E	
LinUCB	**0.1016**	0.0946	0.4580	**0.4646**	0.0956	0.0912	**0.4440**	**0.4374**	0.2734
BerlinUCB	0.0986	**0.0948**	0.4814	0.4480	0.0988	0.0926	0.3980	0.3898	0.2628
B-Kmeans	0.1012	0.0944	0.3964	0.2898	0.1048	0.0854	0.3200	0.1758	0.1960
B-KNN	0.1012	0.0946	0.4638	0.4134	**0.1626**	**0.1576**	0.4188	0.3998	**0.2765**
B-GMM	0.1010	0.0684	**0.5494**	0.2208	0.0926	0.0694	0.3992	0.2002	0.2126

Table 4: **Accuracy:** Nonstationary rewards with varying episodic rewards

	shuffled class labels				multi-task setting		average
	MNIST-F	MNIST-E	Warfarin-F	Warfarin-E	MNIST/Warfarin-F	MNIST/Warfarin-E	
LinUCB	0.1080	0.0974	**0.6464**	**0.6348**	0.3496	0.3442	0.3634
BerlinUCB	0.1126	0.1036	0.6116	0.6080	0.3136	0.3071	0.3428
B-Kmeans	0.2376	0.2566	0.5262	0.5152	0.3801	0.3690	0.3808
B-KNN	**0.2574**	**0.2582**	0.6278	0.6026	**0.3833**	**0.4041**	**0.4222**
B-GMM	0.0958	0.0664	0.5488	0.4038	0.2052	0.2375	0.2596

Application: Online Speaker Recognition

An interactive speaker recognition system learns through the user's feedback on whether the system is correct or not. However, this reward is very sporadic as user don't monitor 24/7.



Evaluation on MiniVox benchmark:

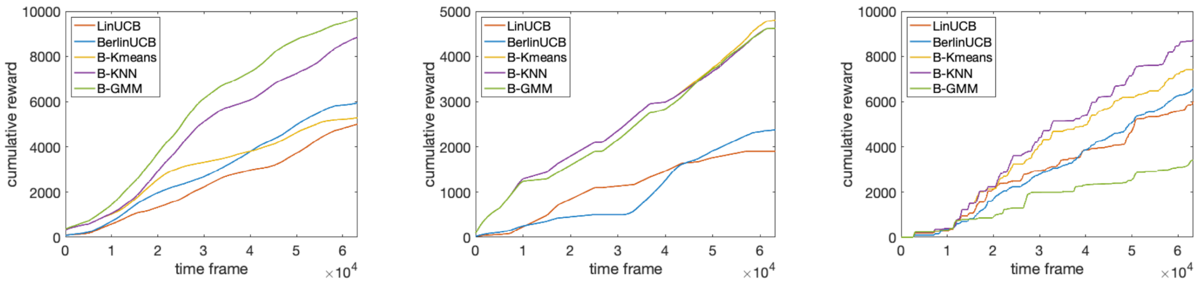


Figure 2: **Cumulative rewards in MiniVox**. (a) p_r=0.1; (b) p_r=0.01; (c) p_r=0.001.

An extended version of this application can be found at:

- Lin & Zhang, "VoiceID on the fly: A Speaker Recognition System that Learns from Scratch" *INTERSPEECH 2020* (demo)
- Lin & Zhang, "Speaker Diarization as a Fully Online Learning Problem in MiniVox" under review in *ICASSP 2021* (arXiv:2006.04376)

Conclusion

A novel problem setting from practical online learning applications

- *Contextual Bandits with Episodic Reward*

A novel solution introducing self-supervision for semi-supervision

- *Background Episodically Rewarded LinUCB (BerlinUCB)*

A novel nonstationary benchmark

- Six synthetic nonstationary settings on context, reward and oracle.

Ongoing Directions

- Improve self-supervision with graph-based methods
- Incorporate online clustering together with the online learning problem
- Theoretical work
- More applications

Several empirical observations

- Updating the representation structure when no reward is revealed improves performance of contextual bandits.
- Adaptive learning with context-dependent clustering modules is much better than learning without self-supervision.

Thank you!

Feel free to contact me if you have any questions.

Check out several of our related works

- Lin et al., “Contextual Bandit with Adaptive Feature Extraction” *ICDMW 2018*
- Lin & Zhang, “VoiceID on the fly: A Speaker Recognition System that Learns from Scratch” *INTERSPEECH 2020*
- Lin & Zhang, “Speaker Diarization as a Fully Online Learning Problem in MiniVox” under review in *ICASSP 2021* (arXiv)

Full paper: https://arxiv.org/abs/2009.08457

Full codes: https://github.com/doerlbh/BerlinUCB