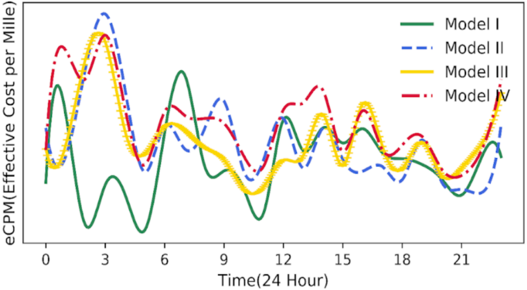
Contextual Active Model Selection

Xuefeng Liu[1] Fangfang Xia[2] Rick L. Stevens[1,2] Yuxin Chen[1]

[1]University of Chicago [2]Argonne National Laboratory

Motivation

- As pre-trained models become increasingly prevalent in a variety of real-world machine learning applications, there is a growing demand for label-efficient approaches for model selection
- No single pre-trained model achieves the best performance for every context
- Model performance depends on the context
- Cost-sensitive to evaluate and access the models / labels
- Online streaming data instead of a pool of data points



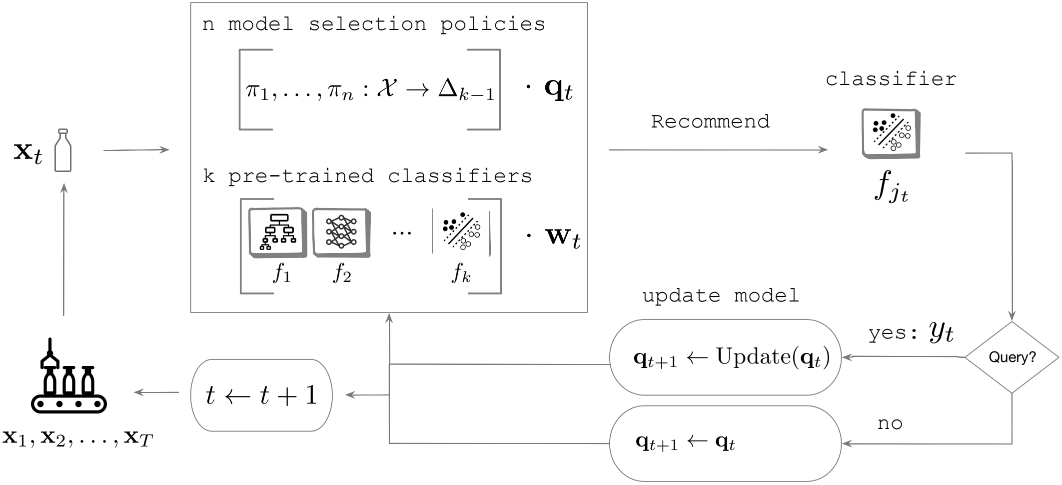
Research question

- How to select data-adaptive models when facing heterogeneous data stream?
- How to make it labeling efficient?

- We want a robust cost-effective online-learning algorithms that
  - effectively identify best model selection policy
  - works under limited labeling resources
  - adaptive to arbitrary data streams

We formally define it as the Contextual Active Model Selection (CAMS) problem and propose a novel algorithm, also named CAMS, to effectively address it.

Learning Protocol



Learning Protocol

Algorithm Contextual Active Model Selection Protocol

Given a set of classifiers $\mathcal{F}$ and model selection policies Π
for $t = 1, 2, ..., T$ **do**
    The learner receives a data instance $\mathbf{x}_t \in \mathcal{X}$ as the context for the current round
    Compute the predicted label $\hat{y}_{t,j} = f_j(\mathbf{x}_t)$ for pre-trained classifier indexed by $j \in [k]$
    The learner identifies a model/classifier f_{j_t} and makes a prediction $\hat{y}_{t,j_t}$ for the instance $\mathbf{x}_t$ based on previous observations.
    if The learner decide to query **then**
        The learner incurs a query cost
        The learner observes true label y_t and receives a (full) 0-1 loss vector $\boldsymbol{\ell}_t = \mathbb{I}_{\{\hat{\mathbf{y}}_t \neq y_t\}}$
        The learner can then use the queried labels to adjust its model selection criterion for future rounds.

Method Overview

Input: Models $\mathcal{F}$, policies Π, #rounds T, budget b
Initialize loss $\tilde{\boldsymbol{L}}_0 \leftarrow 0$; query cost $C_0 \leftarrow 0$
Set $\Pi^* \leftarrow \Pi \cup \{\pi_1^{\text{const}}, \ldots, \pi_k^{\text{const}}\}$
for $t = 1, 2, ..., T$ **do**
  Receive $\boldsymbol{x}_t$
  $\eta_t \leftarrow \text{SetRate}(t, \boldsymbol{x}_t, |\Pi^*|)$
  Set $q_{t,i} \propto \exp\left(-\eta_t \tilde{L}_{t-1,i}\right) \; \forall i \in |\Pi^*|$
  $j_t \leftarrow \text{Recommend}(\boldsymbol{x}_t, \boldsymbol{q}_t)$
  Output $\hat{y}_{t,j_t} \sim f_{t,j_t}$ as the prediction for $\boldsymbol{x}_t$
  Compute $z_t = \max\{\delta_0^t, \mathfrak{E}(\hat{\boldsymbol{y}}_t, \boldsymbol{w}_t)\}$
  Sample $U_t \sim \text{Ber}(z_t)$
  if $U_t = 1$ and $C_t \leq b$ **then**
    Query the label y_t
    $C_t \leftarrow C_{t-1} + 1$
    Compute $\boldsymbol{\ell}_t$: $\ell_{t,j} = \mathbb{I}\{\hat{y}_{t,j} \neq y_t\}, \forall j \in [|\mathcal{F}|]$
    Estimate model loss: $\hat{\ell}_{t,j} = \frac{\ell_{t,j}}{z_t}, \forall j \in [|\mathcal{F}|]$
    Update $\tilde{\boldsymbol{\ell}}_t$: $\tilde{\ell}_{t,i} \leftarrow \langle \pi_i(\boldsymbol{x}_t), \hat{\ell}_{t,j} \rangle, \forall i \in [|\Pi^*|]$
    $\tilde{\boldsymbol{L}}_t = \tilde{\boldsymbol{L}}_{t-1} + \tilde{\boldsymbol{\ell}}_t$
  else
    $\tilde{\boldsymbol{L}}_t = \tilde{\boldsymbol{L}}_{t-1}$
    $C_t \leftarrow C_{t-1}$
  end if
end for

Contextual Model Selection
Active Queries
Model Updates

procedure SetRate$(t, \boldsymbol{x}_t, m)$
  if STOCHASTIC **then**
    $\eta_t = \sqrt{\frac{\ln m}{t}}$
  end if
  if ADVERSARIAL **then**
    Set ρ_t as in adversarial setting section
    $\eta_t = \sqrt{\frac{1}{\sqrt{t}} + \frac{\rho_t}{c^2 \ln c}} \cdot \sqrt{\frac{\ln m}{T}}$
  end if
  return η_t
end procedure

procedure Recommend$(\boldsymbol{x}_t, \boldsymbol{q}_t)$
  if STOCHASTIC **then**
    $\boldsymbol{w}_t = \sum_{i \in |\Pi^*|} q_{t,i} \pi_i(\boldsymbol{x}_t)$
    $j_t \leftarrow \text{maxind}(\boldsymbol{w}_t)$
  end if
  if ADVERSARIAL **then**
    $i_t \sim \boldsymbol{q}_t$
    $j_t \sim \pi_{i_t}(\boldsymbol{x}_t)$
  end if
  return j_t
end procedure

We denote by $\bar{\ell}_t^y := \langle \mathbf{w}_t, \mathbb{I}\{\hat{\boldsymbol{y}}_t \neq y\} \rangle$ as the expected loss if the true label is y, where $\mathbf{w}_t = \pi_{\text{maxind}(\boldsymbol{q}_t)}(\boldsymbol{x}_t)$ and $\text{maxind}(\mathbf{w}) := \arg\max_{j: w_j \in \mathbf{w}} w_j$.

Comparison against Related Work

Algorithm / Setup	**Online bagging** bagging	**Hedge** online learning	**EXP3** bandit	**EXP4** contextual bandits	**Query by Committee** active learning	**ModelPicker** model selection	CAMS (ours)
model selection	×	✓	✓	✓	×	✓	✓
full-information	✓	✓	×	×	✓	✓	✓
active queries	×	×	×	×	✓	✓	✓
context-aware	×	×	×	✓	×	×	✓

Theoretical Guarantees

- **Pseudo-regret** for stochastic setting

$$\overline{\mathcal{R}}_T\left(\mathcal{A}\right) = \mathbb{E}[L_T^{\mathcal{A}}] - T \min_{i \in [|\Pi^*|]} \mu_i, \quad (1)$$

where μ_i represents the expected loss of policy i if recommending the most probable model $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathbf{x}_t, y_t}\left[\ell_{t,\text{maxind}(\pi_i(\mathbf{x}_t))}\right]$.

- **Expected regret** for adversarial setting

$$\mathcal{R}_T\left(\mathcal{A}\right) = \mathbb{E}[L_T^{\mathcal{A}}] - \min_{i \in [|\Pi^*|]} \sum_{t=1}^{T} \tilde{\ell}_{t,i}, \quad (2)$$

where $\tilde{\ell}_{t,i}$ represents the expected loss of policy i if randomizing the model recommendation at t, $\tilde{\ell}_{t,i} := \langle \pi_i\left(\mathbf{x}_t\right), \boldsymbol{\ell}_t \rangle$

Theoretical Guarantees

- **Query complexity** and **tight regret bound** under
  - Stochastic data streams
  - Adversarial data streams
  - Finite policy / model classes

Algorithm	Regret	Query Complexity
Exp3	$2\sqrt{Tk \log k}$	–
Exp4	$\sqrt{2Tk \log G}$	–
Model Picker$_{\text{stochastic}}$	$62 \max_i \Delta_i k / \left(\lambda^2 \log k\right)$ $\lambda = \min_{j \in [k] \setminus \{i^*\}} \Delta_j^2 / \theta_j$	$\sqrt{2T \log k}(1 + 4\frac{c}{\Delta})$
Model Picker$_{\text{adversarial}}$	$2\sqrt{2T \log k}$	$5\sqrt{T \log k} + 2L_{T,*}$
CAMS$_{\text{stochastic}}$	$\left(\frac{\ln \frac{\lvert\Pi^*\rvert - 1}{\gamma} + \sqrt{\ln \lvert\Pi^*\rvert \cdot 2b^2 \ln \frac{2}{\delta}}}{\sqrt{\ln \lvert\Pi^*\rvert} \Delta} \right)^2$	$\left(\left(\frac{\ln \frac{\lvert\Pi^*\rvert - 1}{\gamma} + \sqrt{\ln \lvert\Pi^*\rvert \cdot 2b^2 \ln \frac{2}{\delta}}}{\sqrt{\ln \lvert\Pi^*\rvert} \Delta} \right)^2 + T\mu_{i^*} \right) \frac{\ln T}{c \ln c}$
CAMS$_{\text{adversarial}}$	$2c\sqrt{\ln c / \max\{\rho_T, \sqrt{1/T}\}} \cdot \sqrt{T \log \lvert\Pi^*\rvert}$	$O\left(\left(\sqrt{\frac{T \log \lvert\Pi^*\rvert}{\max\{\rho_T, \sqrt{1/T}\}}} + \tilde{L}_{T,*} \right) (\ln T) \right)$

Experiments

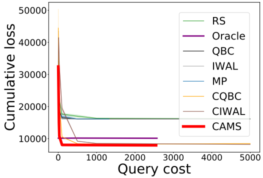
dataset	classification	total instances	test set	stream size	classifier	policy
CIFAR10	10	60000	10000	10000	80	85
DRIFT	6	13910	3060	3000	10	11
VERTEBRAL	3	310	127	80	6	17
HIV	2	40000	4113	4000	4	20
CovType	55	580000	100000	100000	6	17

- Baselines

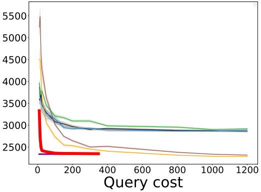
  context-free: Random Sampling (RS, query the instance label with fixed probability), Query by Committee (QBC, committee-based sampling), Importance Weighted Active Learning (IWAL, Calculate query probability based on labeling disagreements of surviving classifiers), Model Picker (MP, employ variance based active sampling)

  contextual: Contextual-QBC (CQBC), Contextual-IWAL (CIWAL), Oracle

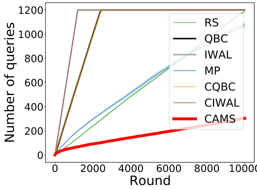
Experiments



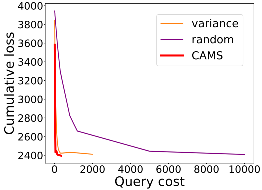
(a) COVTYPE



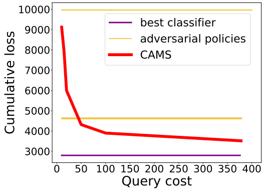
(b) CIFAR10



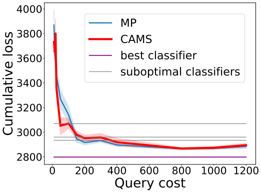
(c) Query complexity study, maintaining label efficiency



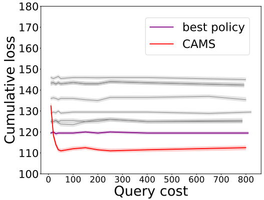
(d) Query strategy ablation



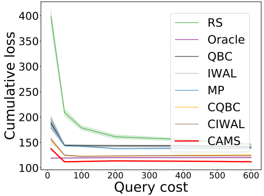
(e) Robustly recover from malicious advices



(f) Reduction to context-free setting when no experts are available



(g) CAMS vs best policy (HIV), effectively converge to the best expert



(h) Adjust probability (HIV), possibly outperforming all the experts including oracle