Asking the Right Questions to the Right Users: Active Learning with Imperfect Oracles

Shayok Chakraborty
Department of Computer Science
Florida State University

AAAI-2020

Background

Imperfect oracles
- Oracles have diverse expertise

Batch mode active learning
- Select a batch of samples instead of single sample

About this work

Samples **selection** and **assignment**

Notation

L: training set
U: unlabeled set
w: model trained on L
Z: number of classes
k: number of oracles
O_i: i-th oracle
C_i: cost of O_i

Sample selection

Informativeness

$$E(x_i) = -\sum_{j=1}^{Z} p_{ij} log p_{ij}$$

Redundancy

$$R(i,j) = \min(0, \cos(x_i, x_j)$$

Oracle selection

Error probability (trained on L)

$$q_{ij}$$

cost

$$C_j = \alpha A_j$$

$\boldsymbol{p(j,i)} = \frac{q_{ij} * C_j}{E(x_i)}, i = 1, \ldots, |U|, j = 1, \ldots, k$

$\boldsymbol{M} \in \{0,1\}^{|U| \times k}$

$\boldsymbol{e}$ -vector of length k with all entry 1

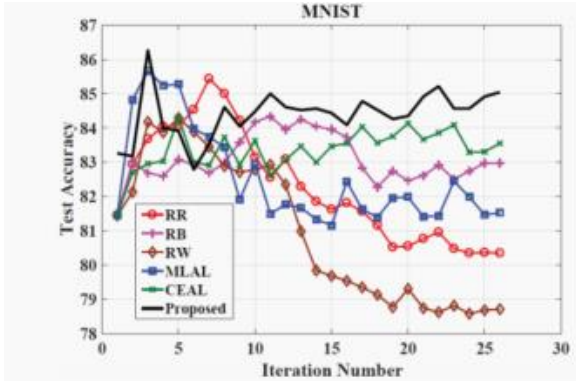
$$\min_{M} \; trace(MP) + \lambda (Me)^T R(Me)$$
$$s.t. \; M_{ij} \in \{0,1\}, \forall i,j$$
$$M_i . e \leq 1, \forall i$$
$$< M, E > = B$$

$$\mathrm{MP} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ 0 & 0 & 0 \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$$

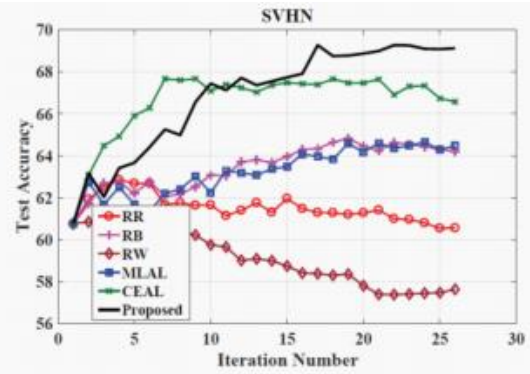
$$(Me)^T R(Me) = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)^T R \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

$$= (r_{11} + r_{13}) + (r_{31} + r_{33})$$

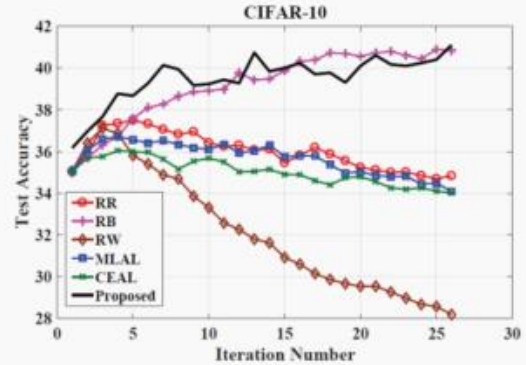
$$\begin{bmatrix} r_{11} & r_{12} & r_{13} \end{bmatrix}$$
$$+$$
$$\begin{bmatrix} r_{31} & r_{32} & r_{33} \end{bmatrix}$$



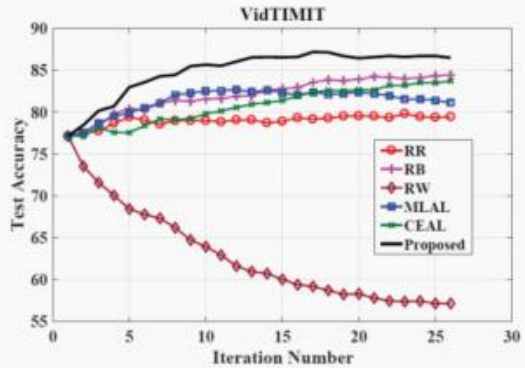
(a) MNIST



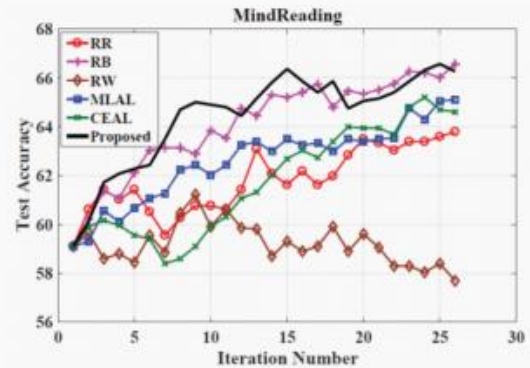
(b) SVHN



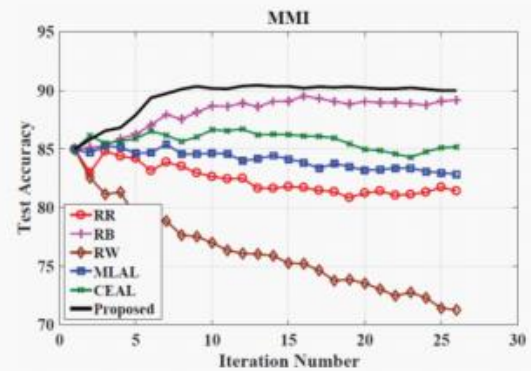
(c) CIFAR



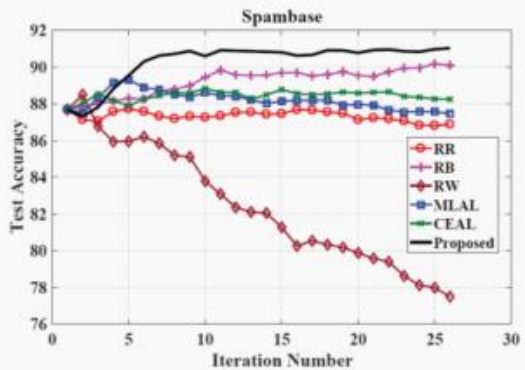
(d) VidTIMIT



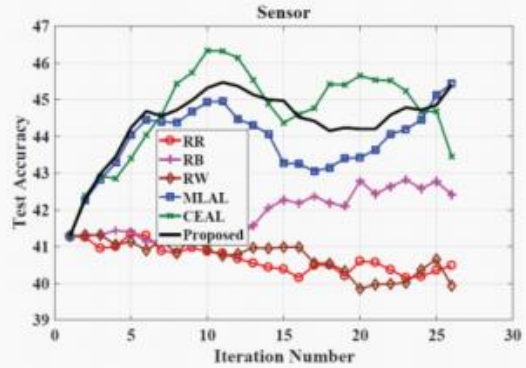
(e) MindReading



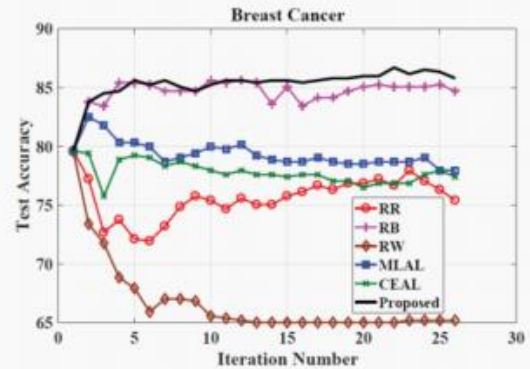
(f) MMI



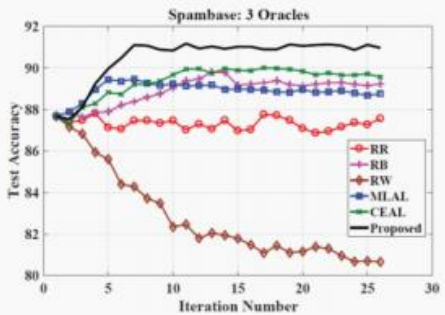
(g) Spambase



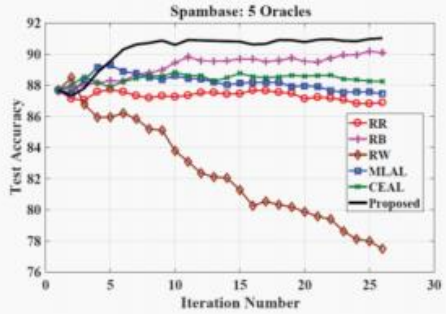
(h) Sensor



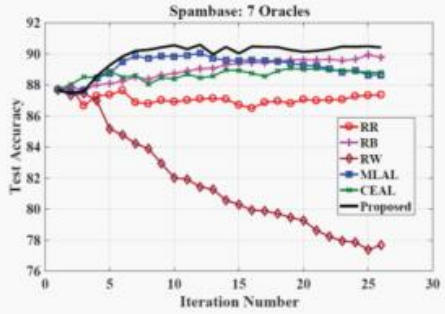
(i) Breast Cancer



(a) 3 Oracles

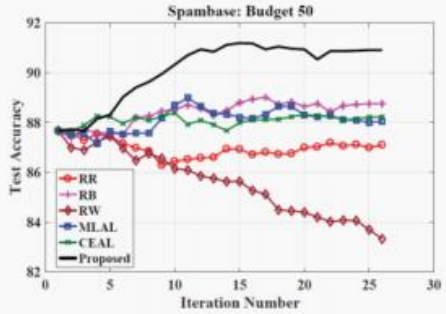


(b) 5 Oracles

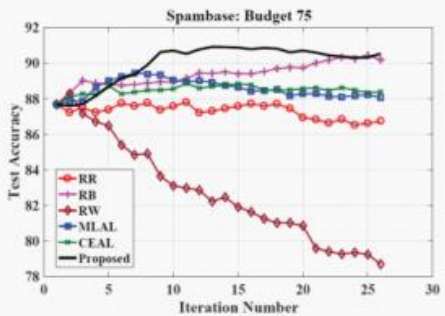


(c) 7 Oracles

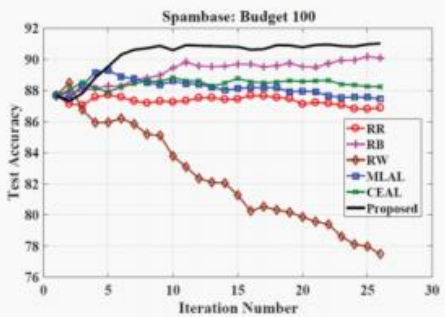
Figure 2: Effect of the number of labeling oracles on the Spambase dataset. Best viewed in color.



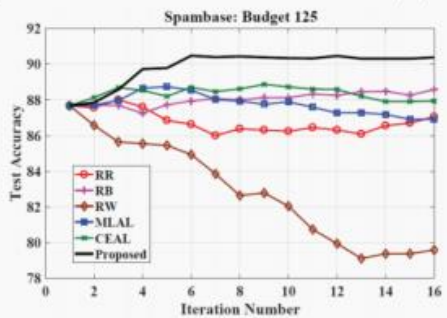
(a) Budget 50



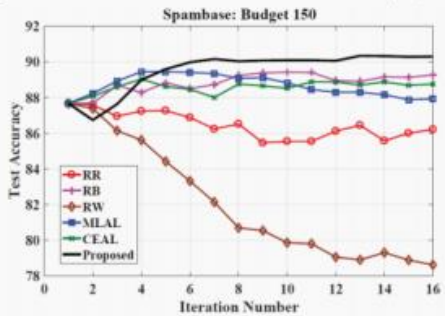
(b) Budget 75



(c) Budget 100



(d) Budget 125



(e) Budget 150

Figure 3: Effect of varying query budgets on the Spambase dataset. Best viewed in color.