Bounding Uncertainty for Active Batch Selection

Hanmo Wang Runwu Zhou YiDong Shen
State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, China
University of Chinese Academy of Sciences

AAAI-2019

Introduction

BMAL: allows the learner to query instance in groups, select both representative and uncertain samples.

The direct approach: directly combines representativeness with uncertainty to form a single objective.

The screening approach: excludes some unlabeled instances about which the classifier is certain, and chooses representative samples among the remaining instances.

Motivation

The two approaches both have shortcomings in the initial stage of BMAL,when the labeled data is scarce:

The direct approach: directly utilizes the output, possibly misled.

The screening approach: screens samples beforehand, may accidentally remove useful instances.

LBC(lower-bounded certainty)

lower-bounded certainty (LBC) :

$$LC(\mathbf{x}_i) = \max(C(\mathbf{x}_i), \epsilon)$$

Combine the representativeness R(S) with LBC score LC(S) and obtain the BMAL framework with LBC:

$$\min_{S \subset U, |S|=b} R(S) + \lambda LC(S)$$

$\epsilon = \min_{i \in U} C(x_i)$

direct approach:

$$\min_{S \subset U, |S|=b} R(S) + \lambda C(S),$$

$\lambda > R(S^*)/v^*$

screening approach:

$$\min_{S \subset U, |S|=b} R(S)$$

$$s.t.\ C(\mathbf{x}_i) \leq \epsilon \ for\ all\ i \in S$$

R(S):representativeness function
C(S):certainty function
$C(x_i) = \max_y P(y|\mathbf{x}_i, \mathbf{w})$

s^* :the global optimizer of screening approach

$v^* = \min_{C(\mathbf{x}_i) > \epsilon} C(\mathbf{x}_i) - \epsilon$ for $i \in U$.

LBC(lower-bounded certainty)

MMD: selects instances that minimizes the difference in distribution between labeled and unlabeled data:

$$R(S) := \mathbf{MMD}[\phi, X_{L\cup S}, X_{U\setminus S}] = \frac{1}{2}\sum_{i\in S}\sum_{j\in S} K_{ij} + \sum_{i\in S} h_i$$

$$h_i = \frac{n_u - b}{n}\sum_{j\in L} K_{ij} - \frac{n_l + b}{n}\sum_{j\in U} K_{ij}$$

The objective function:

$$\min_{S\subset U,|S|=b} f(S) = \frac{1}{2}\sum_{i\in S}\sum_{j\in S} K_{ij} + \sum_{i\in S} h_i + \lambda\sum_{i\in S} LC(\mathbf{x}_i)$$

Random Greedy Solver:

The increase of the objective function f(·) after adding index e to set S can be formulated as :

$$f_S(e) = f(S\cup\{e\}) - f(S) = \sum_{i\in S} K_{ie} + \frac{1}{2}K_{ee} + h_e + \lambda LC(\mathbf{x}_e)$$

After selecting another index e', the increase becomes :

$$f_{S\setminus e'}(e) = \sum_{i\in S\setminus e'} K_{ie} + \frac{1}{2}K_{ee} + h_e + \lambda LC(\mathbf{x}_e) = f_S(e) - K_{ee'}$$

LBC(lower-bounded certainty)

Algorithm 1 summarizes the random greedy algorithm, where ψ and ψ' correspond to f_S and $f_{S\backslash e'}$ respectively.

Algorithm 1 RandGreedy(U,b)% select b instances from U

Require: h, kernel K, batch size b, unlabeled index set U
Ensure: A solution S

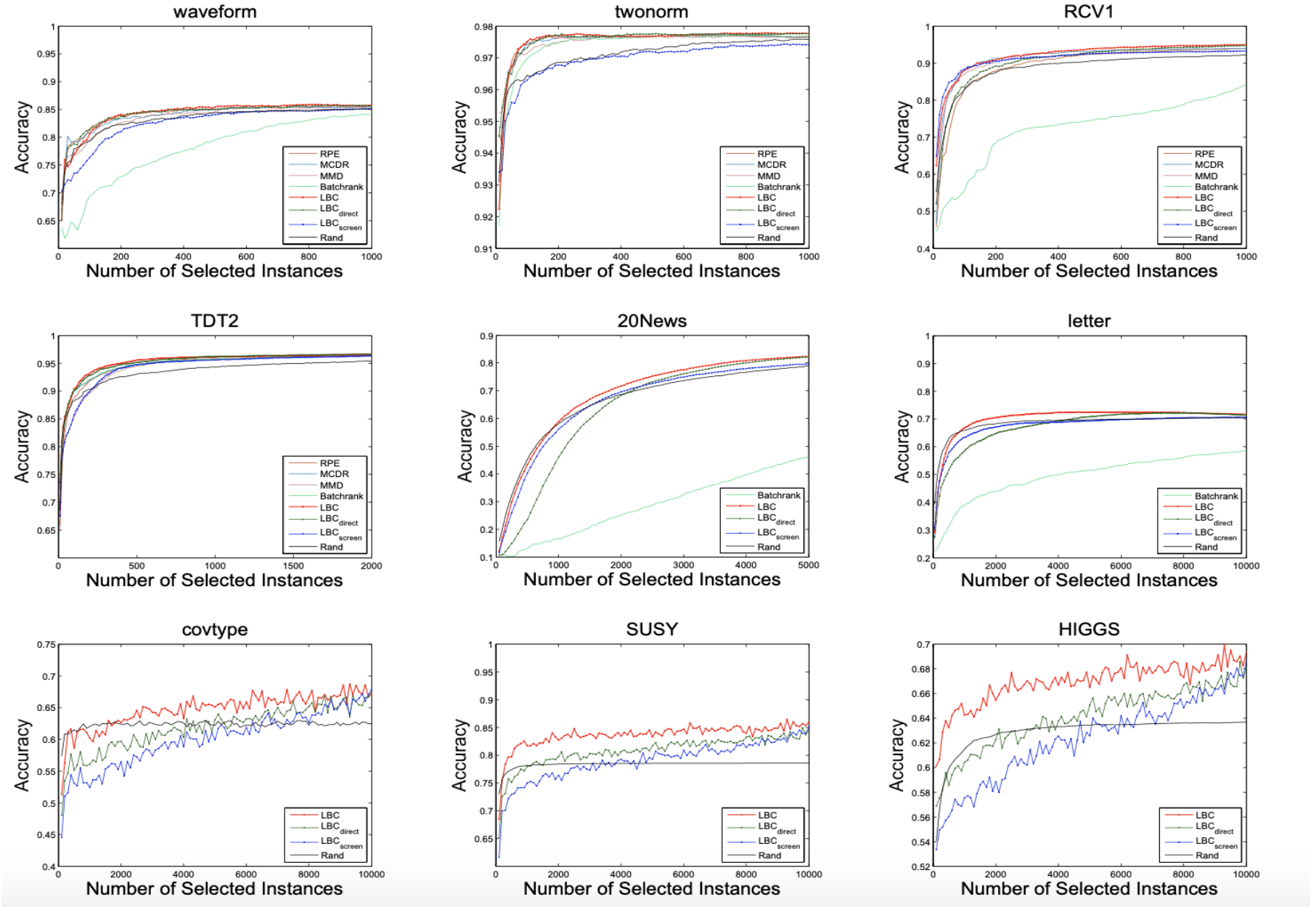
$S \leftarrow \emptyset$
$\psi_e = \frac{1}{2}K_{ee} + h_e + \lambda LC(\mathbf{x}_e)$, for all $e \in U$
for i=1 to b **do**
  Let M be the set of b indexes in $U\backslash S$ with b smallest ψ value
  Randomly select one index e' from M
  $S \leftarrow S \cup \{e'\}$
  $\psi'_e \leftarrow \psi_e - K_{ee'}$, for all $e \in U\backslash S$
  $\psi_e \leftarrow \psi'_e$, for all $e \in U\backslash S$
end for
return S

Algorithm 2 BMAL based on LBC

Require: batch size b, labeled index set L, unlabeled index set U, and data matrix X, batch number t
Ensure: A solution S of size b

if t=1 **then**
  Construct Kernel matrix K from X
  Initialize coefficient $\mathbf{h}^1$ using Eq. (15)
end if
$S \leftarrow$ RandGreedy(U, b)
Update coefficient $\mathbf{h}^{t+1}$ with $\mathbf{h}^t$ using Eq. (15) and Eq. (16)
return S

Experiment


waveform
twonorm
RCV1
TDT2
20News
letter
covtype
SUSY
HIGGS
Accuracy
Number of Selected Instances
RPE
MCDR
MMD
Batchrank
LBC
LBC direct
LBC screen
Rand

Experiment

	Win/Loss(%) of LBC vs. the following						
Dataset	Rand	MMD	RPE	Batchrank	MCDR	direct	screen
ORL	74/0	53/0	53/0	0/0	53/0	0/0	0/0
COIL20	86/0	90/0	54/0	44/0	71/0	20/0	21/0
segmetation	96/0	84/2	98/0	81/0	66/0	11/0	91/0
WEBACE	91/0	89/0	11/1	20/0	52/3	12/1	7/0
Reuters	98/1	99/0	66/0	47/0	98/0	0/0	1/0
USPS	96/1	84/0	32/0	100/0	64/0	9/0	82/4
waveform	85/0	76/0	3/4	99/0	13/4	0/4	90/0
twonorm	78/0	10/2	0/0	11/0	1/0	0/1	97/0
RCV1	99/0	90/0	100/0	100/0	58/0	91/0	78/5
TDT2	97/0	53/0	40/0	44/0	53/0	9/0	72/0
20News	80/14	NA	NA	100/0	NA	92/0	98/0
letter	90/6	NA	NA	100/0	NA	74/0	98/0
covtype	32/2	MLE	MLE	MLE	MLE	3/0	37/0
SUSY	98/1	MLE	MLE	MLE	MLE	89/0	98/0
HIGGS	100/0	MLE	MLE	MLE	MLE	2/0	11/0

Table 3: the win/loss(%) of *LBC* against BMAL baselines using paired t-test with a 95% significant level

Thanks