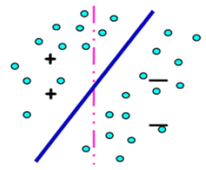
Instance Selection in Semi-Supervised Learning

Yuanyuan Guo, Harry Zhang, and Xiaobo Liu



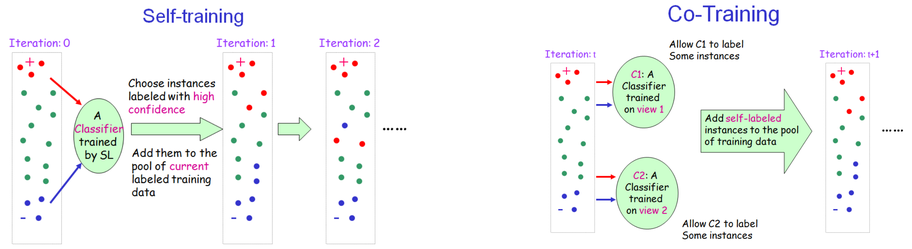
Semi-Supervised Learning (SSL)

- In many real-world applications:
  - Labeled data (*L*): scarce; expensive/difficult to collect.
  - Unlabeled data (*U*): abundant; relatively easy to obtain.
- SSL uses labeled data and unlabeled data to learn hypotheses.

Usage	Supervised learning	Semi-supervised learning	Unsupervised learning
{(x,y)} labeled data	Yes	Yes	No
{x} unlabeled data	No	Yes	Yes

Motivation

- Selecting the most confident unlabeled instances (**"*CF*"**) is a common instance selection method in two standard SSL methods: self-training and co-training.



- Shortcomings of "*CF*" :
  - Label noise may be added to the training set.
  - It is not necessarily superior to that of randomly selecting unlabeled instances.
- Our pervious work showed that the original labeled instances are more reliable than the self-labeled instances that are labeled by the classifier.

Results on 26 benchmark UCI datasets

Accuracy	Self-training		Co-training	
	CF	*ISBOLD*	*CF*	*ISBOLD*
Mean	71.80	74.61	70.34	73.71
Paired t-test		6/20/0		7/19/0

AUC	Self-training		Co-training	
	CF	*ISBOLD*	*CF*	*ISBOLD*
Mean	80.57	83.12	78.56	81.74
Paired t-test		9/17/0		8/18/0

Table 2: AUC of **CF** vs **ISBOLD** in self-training and co-training

(a) self-training

Dataset	CF	ISBOLD
balance-scale	61.37	66.68
breast-cancer	63.98	63.48
breast-w	99.07	99.08
colic	79.24	78.43
colic.ORIG	51.62	58.49
credit-a	86.81	86.79
credit-g	56.56	65.24 v
diabetes	78.03	76.36
heart-c	83.97	83.92
heart-h	83.74	83.74
heart-statlog	88.93	88.64
hepatitis	83.02	80.99
ionosphere	86.86	86.68
iris	98.33	98.29
kr-vs-kp	74.65	89.03 v
labor	96.59	96.72
letter	86.09	93.08 v
mushroom	98.04	98.81 v
segment	90.86	95.24 v
sick	91.51	93.96
sonar	58.64	62.21
splice	94.40	96.23 v
vehicle	59.63	66.95 v
vote	96.31	96.52
vowel	57.65	64.49 v
waveform-5000	88.85	90.96 v
mean	80.57	83.12
w/t/l		9/17/0

(b) co-training

Dataset	CF	ISBOLD
balance-scale	60.44	65.34
breast-cancer	63.51	64.37
breast-w	99.22	99.19
colic	78.99	79.08
colic.ORIG	49.62	55.82
credit-a	88.05	86.35
credit-g	55.33	61.62
diabetes	72.61	74.95
heart-c	84.02	83.80
heart-h	83.77	83.50
heart-statlog	90.03	88.03
hepatitis	78.38	73.19
ionosphere	87.89	88.92
iris	93.21	92.27
kr-vs-kp	66.86	86.39 v
labor	87.76	85.18
letter	82.98	92.57 v
mushroom	97.89	98.75 v
segment	87.93	94.82 v
sick	87.74	93.83
sonar	59.59	62.93
splice	88.65	94.87 v
vehicle	59.56	67.09 v
vote	96.31	96.46
vowel	57.97	66.44 v
waveform-5000	84.22	89.54 v
mean	78.56	81.74
w/t/l		8/18/0

Table 1: Accuracy of **CF** vs **ISBOLD** in self-training and co-training

(a) self-training

Dataset	CF	ISBOLD
balance-scale	59.52	66.21
breast-cancer	65.09	65.61
breast-w	96.67	96.34
colic	74.54	75.38
colic.ORIG	55.05	60.57
credit-a	80.68	80.78
credit-g	60.62	66.03 v
diabetes	70.55	70.53
heart-c	81.55	81.15
heart-h	83.06	82.41
heart-statlog	81.37	80.74
hepatitis	79.70	78.34
ionosphere	80.97	79.86
iris	90.31	90.05
kr-vs-kp	67.26	80.07 v
labor	88.26	87.92
letter	40.38	57.39 v
mushroom	91.90	92.57 v
segment	63.49	72.88 v
sick	91.54	94.15
sonar	55.72	57.93
splice	82.05	85.48 v
vehicle	41.79	48.35
vote	87.89	88.53
vowel	18.75	21.78
waveform-5000	77.98	78.87
mean	71.80	74.61
w/t/l		6/20/0

(b) co-training

Dataset	CF	ISBOLD
balance-scale	59.10	67.17
breast-cancer	70.41	71.00
breast-w	96.85	96.47
colic	76.60	75.76
colic.ORIG	55.19	62.04
credit-a	81.36	79.67
credit-g	63.04	67.72 v
diabetes	67.51	69.58
heart-c	82.77	80.13
heart-h	81.46	78.60
heart-statlog	82.03	80.30
hepatitis	81.04	80.21
ionosphere	81.50	83.08
iris	80.79	78.98
kr-vs-kp	59.22	77.36 v
labor	77.21	78.43
letter	36.67	56.05 v
mushroom	91.74	92.38 v
segment	61.49	71.64 v
sick	93.40	93.56
sonar	55.43	58.08
splice	73.91	82.63 v
vehicle	41.57	47.86
vote	88.21	88.60
vowel	18.83	23.36
waveform-5000	71.61	75.91 v
mean	70.34	73.71
w/t/l		7/19/0

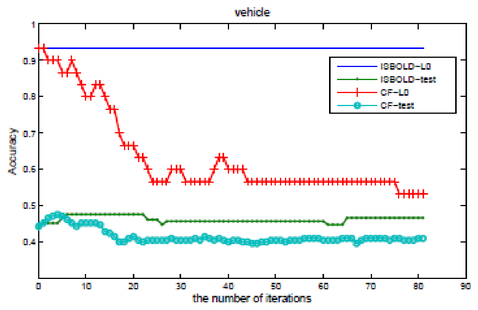
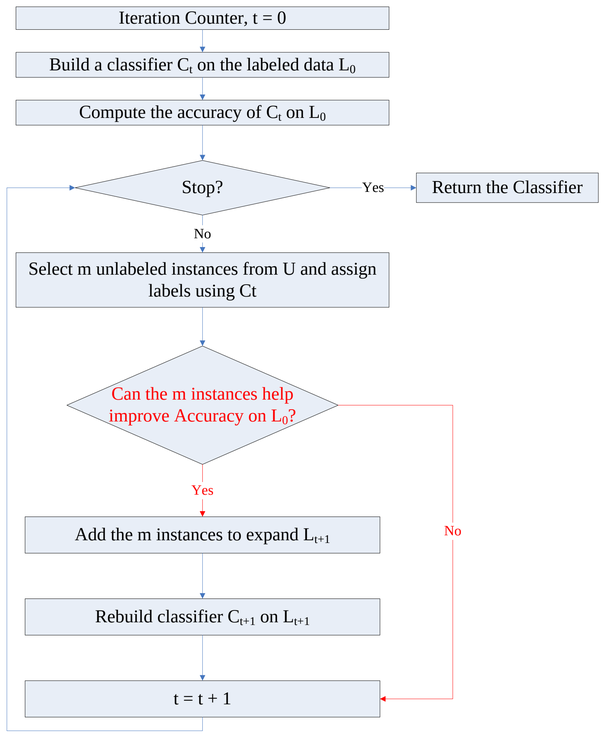


Fig. 3: Learning curves on the *vehicle* dataset

Method

Proposed solution ("***ISBOLD***") : an instance selection method based on the original labeled data



- **Steps**: In each iteration, after the selection of the most confident unlabeled instances,
  a. Compute the accuracy of the current classifier on the original labeled;
  b. Check whether the accuracy is lower than that in last iteration;
  c. If so, discard the selected instances; Otherwise, add them to the training set in the next iteration.
- Reasons:
  - Performance on the original labeled data reflects the performance on the future testing set.
  - It prevents adding unlabeled instances that will possibly degrade the performance.
- Evaluation:
  - 4-fold stratified cross-validation (10 runs), 26 UCI datasets
  - $|L|:|U| = 5\% : 95\%$
  - Measurements: Average accuracy and AUC; Paired t-test
  - Implemented in Weka

Potential real-word applications

- Text mining (e.g., web page classification), Natural language processing (e.g., information extraction)
- Online data stream learning
- Bioinformatics, such as gene expression data analysis
- Medical application (improving computer-aided diagnosis using undiagnosed samples)
- Computer Vision (such as face detection, object detection and object tracking)
- Image classification, traffic classification, intrusion detection, spam detection, and more ……