DMKD’ 17

01 Active Learning: an Empirical Study of Common Baselines

Neurocomputing’ 19

02 Empirical Investigation of Active Learning Strategies

EMNLP-IJCNLP’ 19

03 Practical Obstacles to Deploying Active Learning

Active Learning: an Empirical Study of Common Baselines

Maria E. Ramirez-Loaiza
E-mail: mramire8@hawk.iit.edu

Manali Sharma
E-mail: msharm11@hawk.iit.edu

Geet Kumar
E-mail: gkumar7@hawk.iit.edu

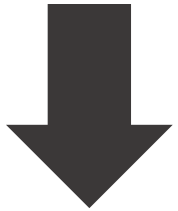
Mustafa Bilgic
E-mail: mbilgic@iit.edu

Illinois Institute of Technology
10 W 31st Street
Chicago, IL, 60616

Motivation

- Most empirical evaluations of AL methods are conducted on very **different datasets with specific metrics and target models**.

- It is **difficult to compare** with different proposed query strategies.



Compare different AL methods under the same setting & report the results

Settings

- Methods

  (1) Uncertainty, (2) QBC, (3) Random

- Metrics

  (1) accuracy, (2) AUC, (3) precision, (4) recall, (5) F1

- Models

  (1) naive Bayes, (2) logistic regression

- Datasets

# of Instances	# of Features	Types of Features	Min. %
20,640	8	Numeric	29%
42,678	1,617	Binary	3.5%
20,722	92	Numeric	37.8%
494,020	41	Numeric + Categorical	16%
20,000	16	Numeric	4%
20,000	16	Numeric	8%
19,466	16,969	Binary	28.5%
50,000	230	Numeric	1.6%
145,252	216	Numeric + Binary	6.2%
61,488	154	Numeric	4.6%

(1) 20 synthesis **binary** datasets with positive class distributions of 50%, 25%, 10%, and 1%

(2) 10 large real-world **binary** classification datasets

Results

Table 5 `UNC` vs. `RND`. Results compare the learning curves of `UNC` against `RND` for `NB` and `LR` classifiers. Win (W), tie (T), and loss (L) counts of statistical significance test (t-test) per measure: AUC, accuracy (ACCU), F_1, precision (PREC), and recall (REC). Results are grouped by synthetic data (SYN) and real-world data (REAL).

DATA-CLF	AUC W/T/L	ACCU W/T/L	F_1 W/T/L	PREC W/T/L	REC W/T/L
SYN-NB	**15**/0/5	**14**/1/5	**15**/1/4	**11**/2/7	**16**/1/3
SYN-LR	**20**/0/0	**17**/1/2	**18**/2/0	**18**/2/0	**18**/2/0
REAL-NB	**5**/0/**5**	**10**/0/0	**10**/0/0	**10**/0/0	2/1/**7**
REAL-LR	4/1/**5**	**7**/2/1	**8**/0/2	**7**/1/2	**8**/1/1

Result 2: **RND was fairly competitive for AUC on real datasets**. RND won significantly over AL on at least five out of ten datasets for AUC for both classifiers.

Results

Table 6 `QBC` vs. `RND`. Results compare the learning curves of `QBC` against `RND` for `NB` and `LR` classifiers. Win (W), tie (T), and loss (L) counts of statistical significance test (t-test) per measure: AUC, accuracy (ACCU), F_1, precision (PREC), and recall (REC). Results are grouped by synthetic data (SYN) and real-world data (REAL).

DATA-CLF	AUC W/T/L	ACCU W/T/L	F_1 W/T/L	PREC W/T/L	REC W/T/L
SYN-NB	**12**/0/8	**13**/3/4	**12**/3/5	8/3/**9**	**13**/2/5
SYN-LR	**15**/0/5	**15**/5/0	**15**/5/0	**13**/6/1	**15**/5/0
REAL-NB	**5**/0/**5**	**10**/0/0	**10**/0/0	**10**/0/0	**5**/0/**5**
REAL-LR	3/1/**6**	**9**/1/0	**8**/0/2	**7**/0/3	**8**/0/2

Result 2: **RND was fairly competitive for AUC on real datasets**. RND won significantly over AL on at least five out of ten datasets for AUC for both classifiers.

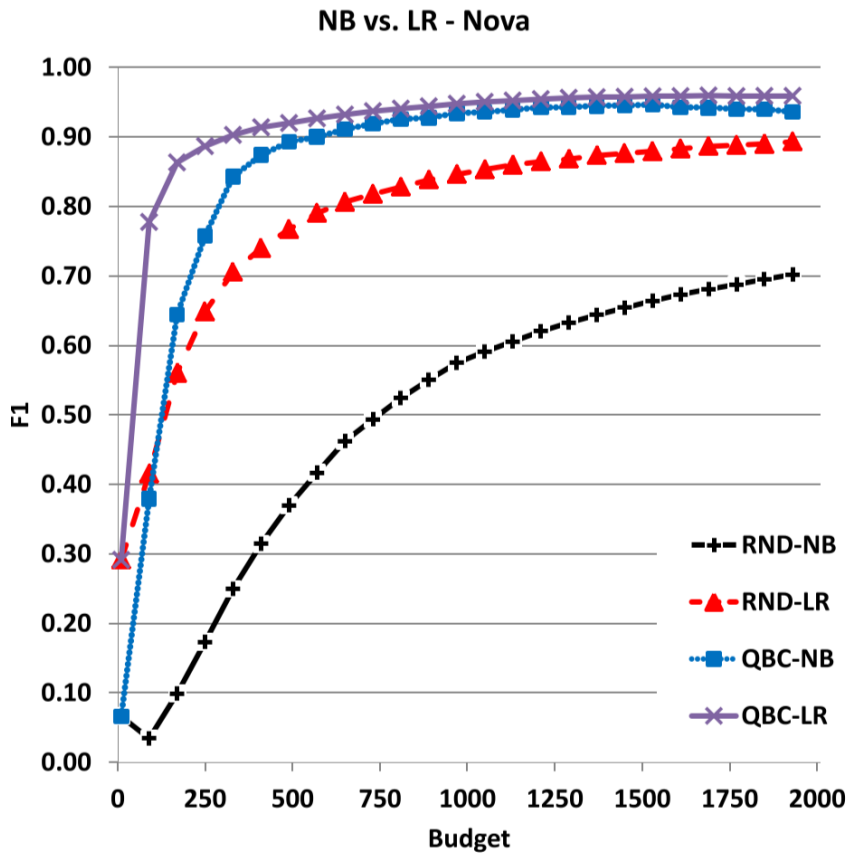
Results

Table 7 `QBC` vs. `UNC`. Results compare the learning curves of `QBC` against `UNC` for `NB` and `LR` classifiers. Win (W), tie (T), and loss (L) counts of statistical significance test (t-test) per measure: AUC, accuracy (ACCU), F_1, precision (PREC), and recall (REC). Results are grouped by synthetic data (SYN) and real-world data (REAL).

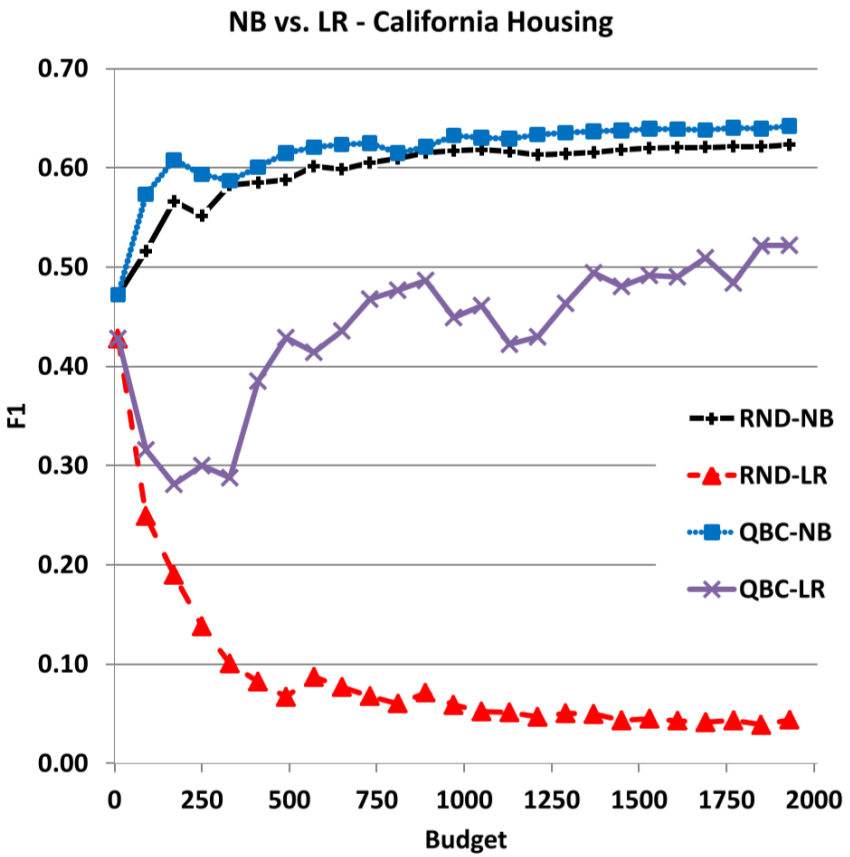
DATA-CLF	AUC W/T/L	ACC. W/T/L	F_1 W/T/L	PREC. W/T/L	REC. W/T/L
SYN-NB	**9**/5/6	**10**/5/5	4/4/**12**	**10**/2/8	2/1/**17**
SYN-LR	3/2/**15**	4/1/**15**	2/3/**15**	5/2/**13**	2/3/**15**
REAL-NB	**5**/0/**5**	2/0/**8**	**5**/0/**5**	0/2/**8**	**8**/0/2
REAL-LR	**5**/1/4	**6**/2/2	3/2/**5**	4/1/**5**	4/0/**6**

Result 5: **Model selection**, which is not trivial for AL, **provides improvements beyond what AL can provide.**

Results



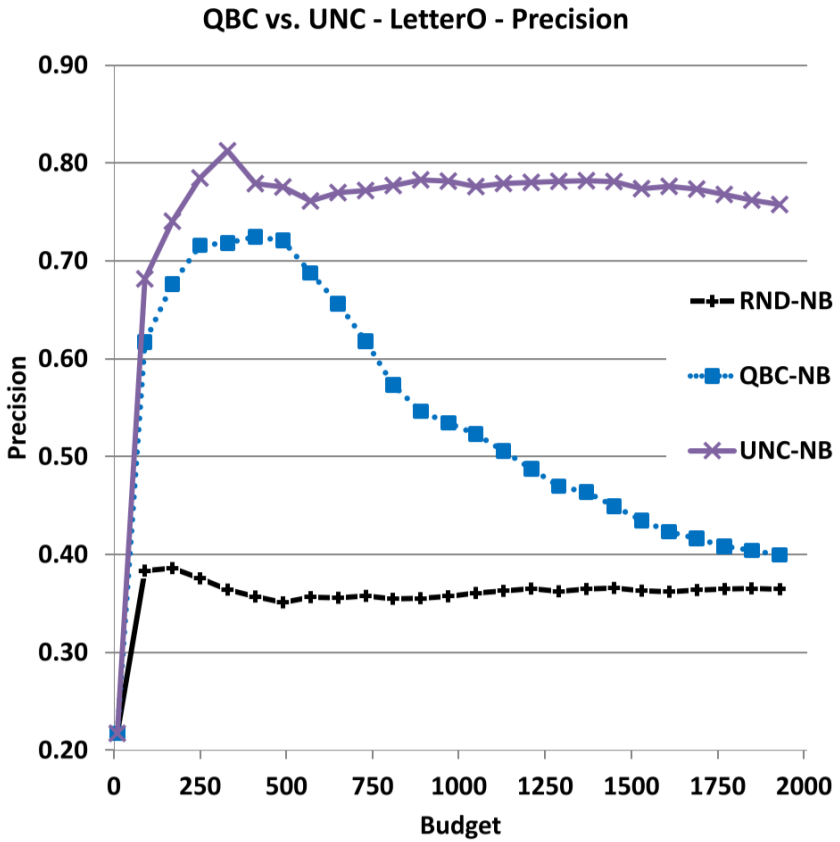
(a)



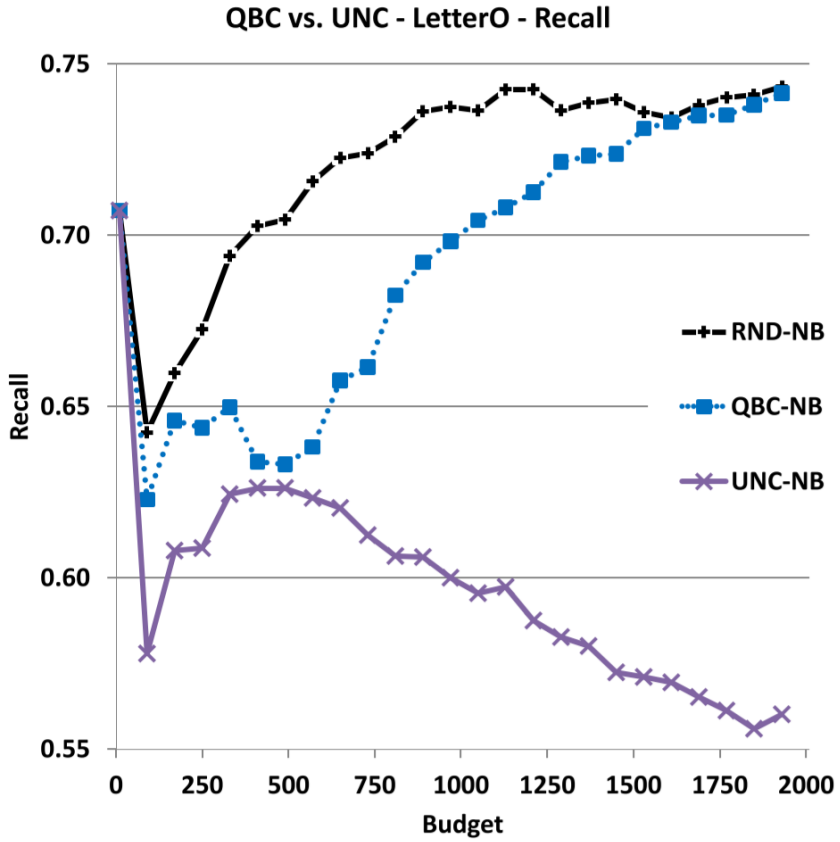
(b)

Result 5: **Model selection**, which is not trivial for AL, **provides improvements beyond what AL can provide.**

Results



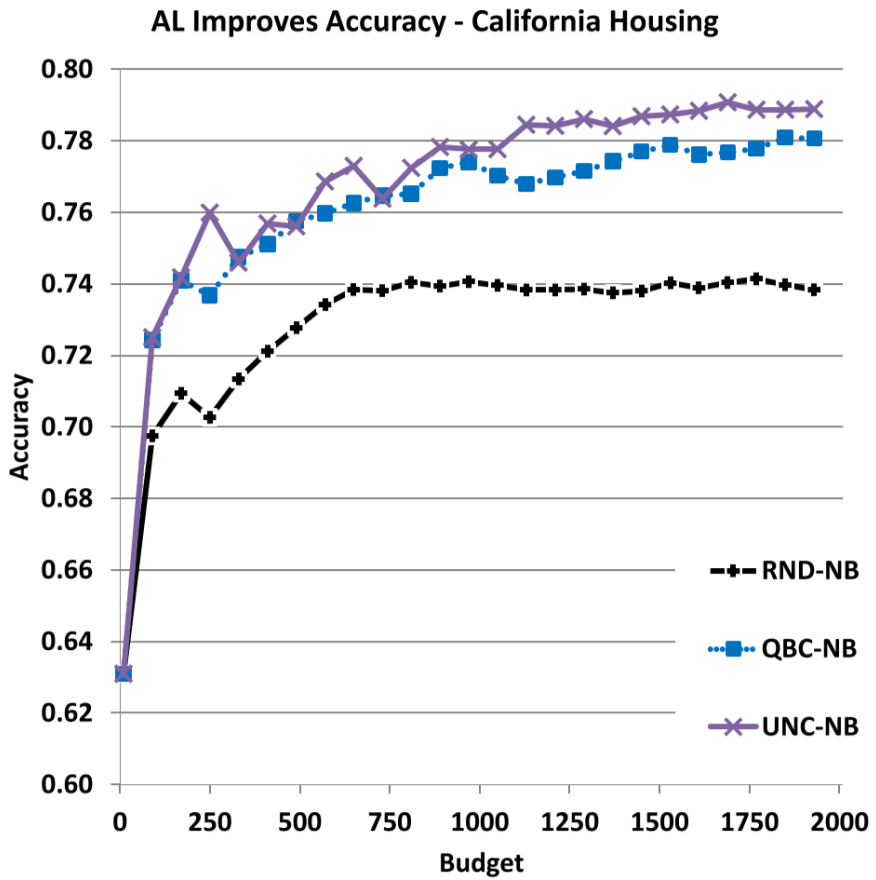
(a)



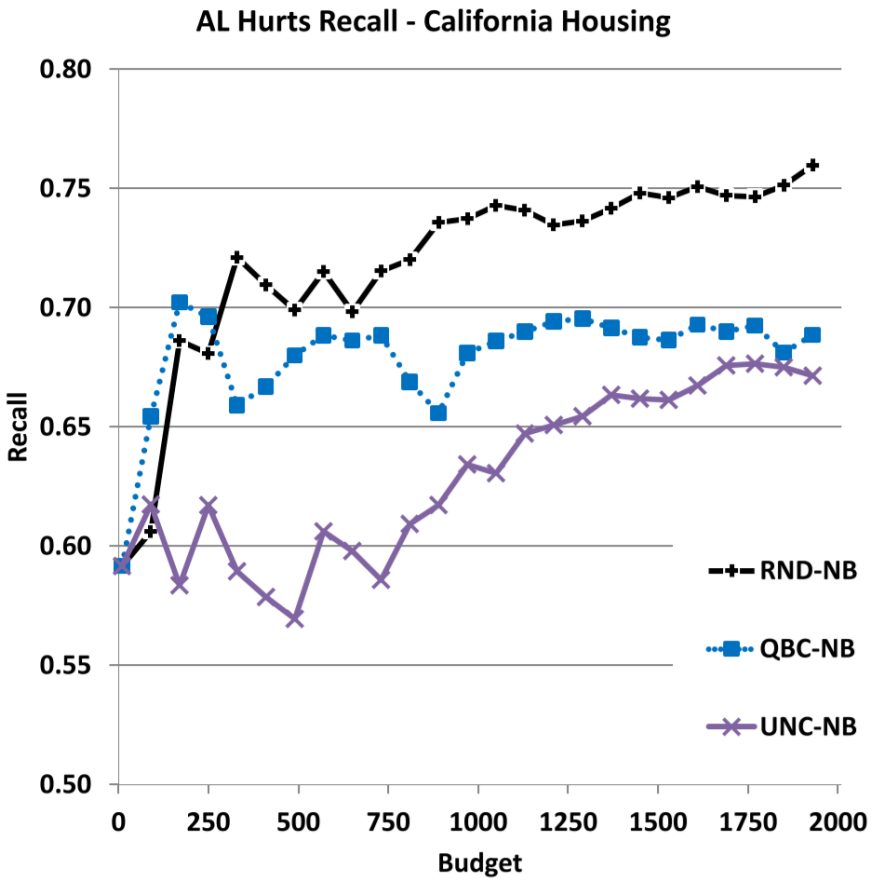
(b)

Result 8: Improvements across the board were rare. **Improvement in one measure often came at the expense of another.** AL often improved accuracy and precision at the expense of recall.

Results



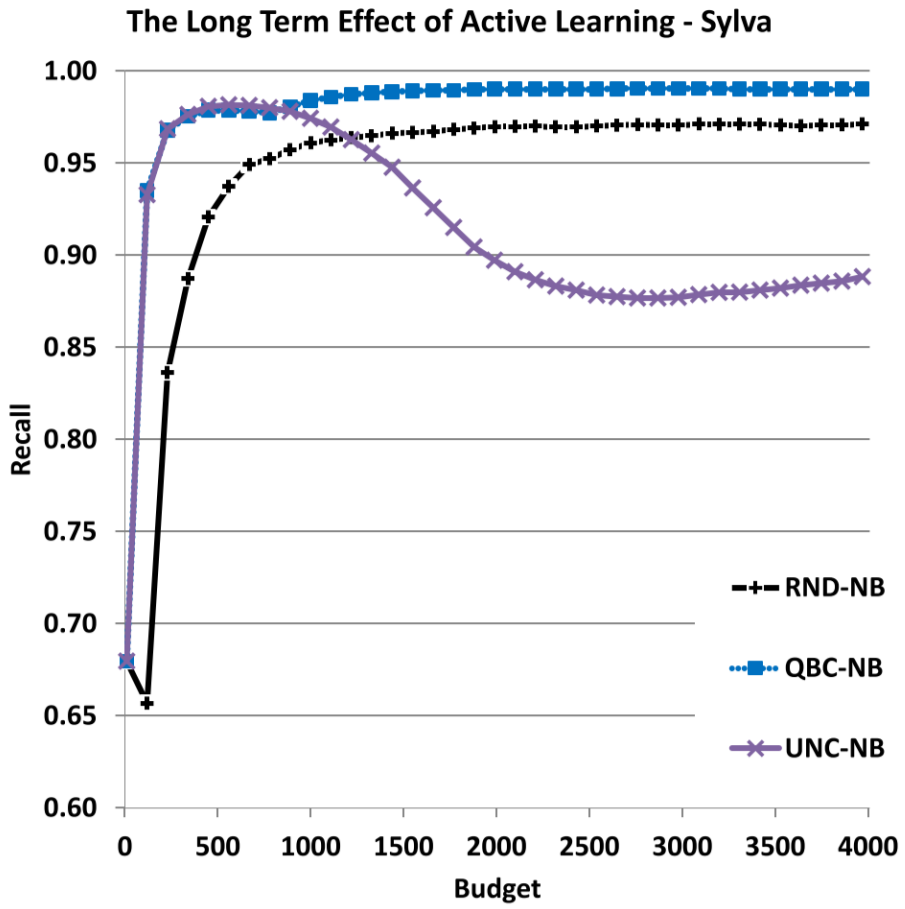
(a)



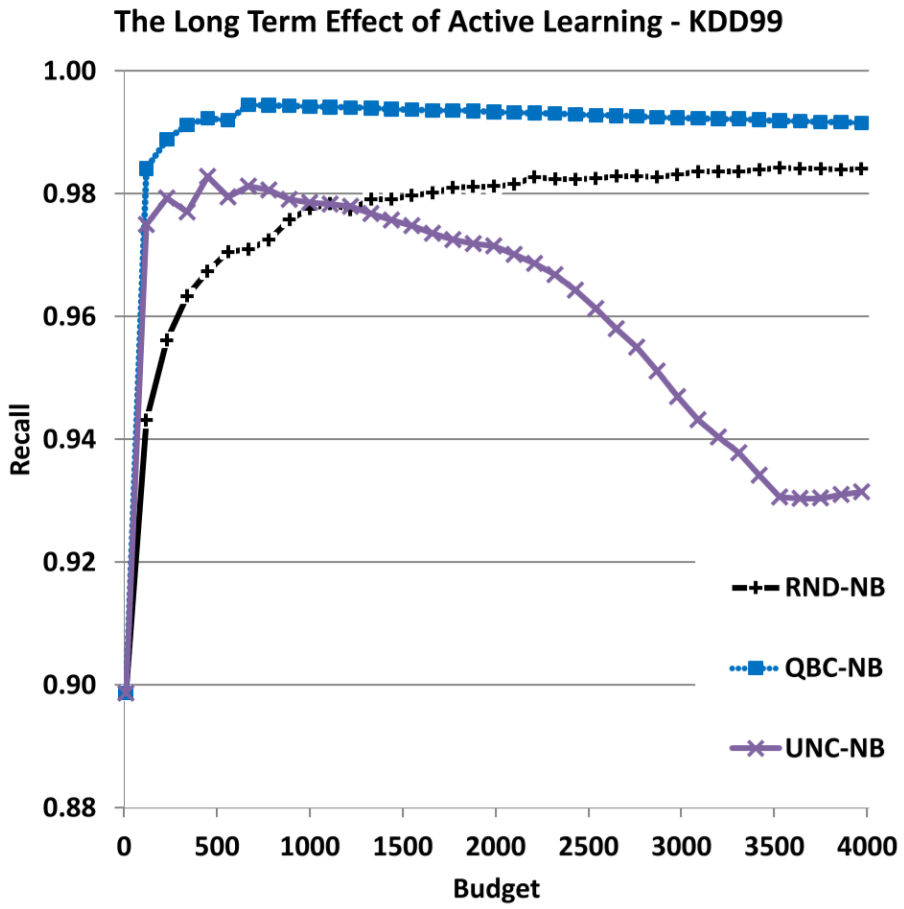
(b)

Result 8: Improvements across the board were rare. **Improvement in one measure often came at the expense of another.** AL often improved accuracy and precision at the expense of recall.

Results



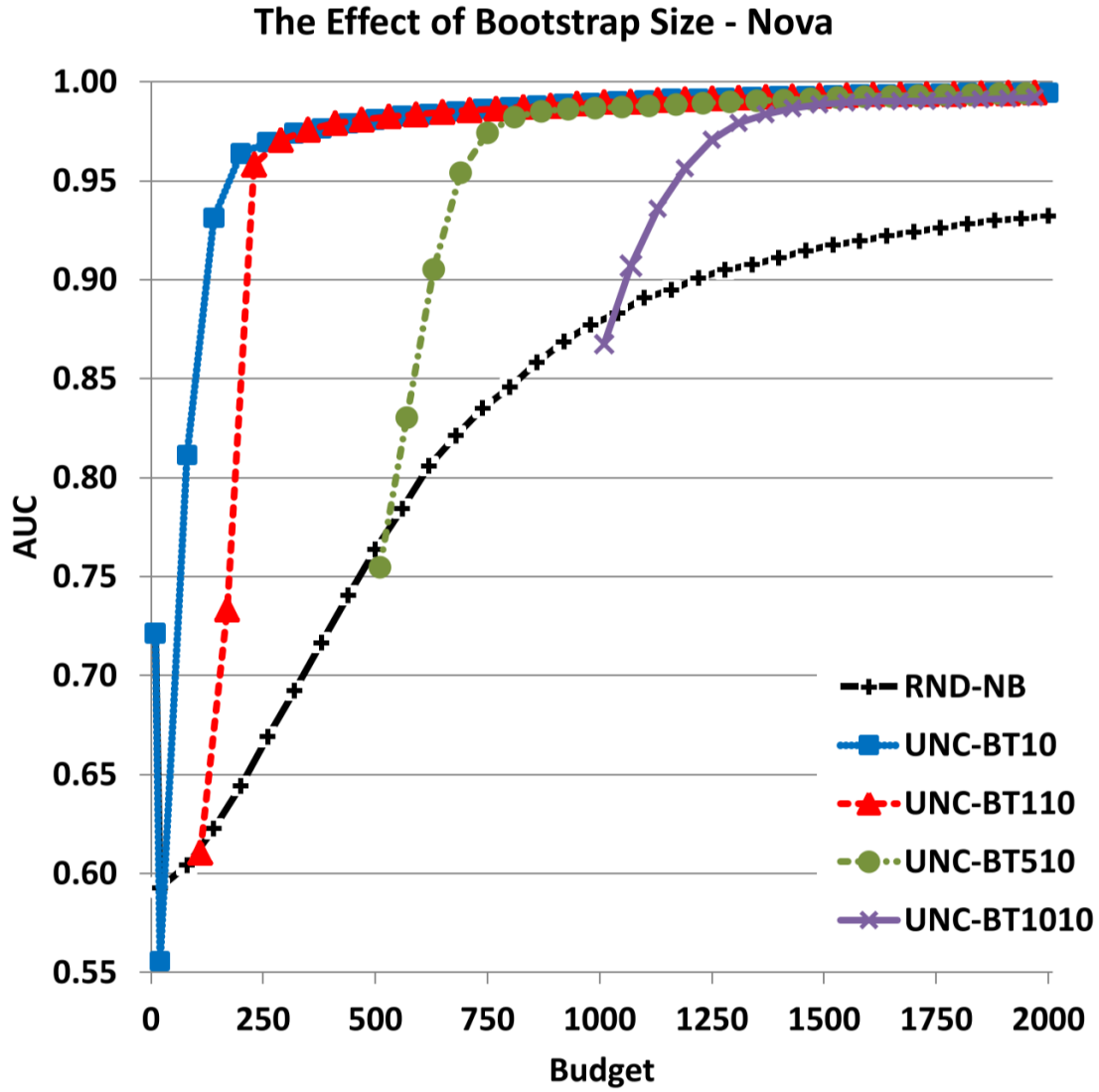
(a)



(b)

Result 10: Continuous labeling with AL can do more harm than good.

Results



Result 12: Using a larger-size initially-labeled data never made a losing AL strategy a winning strategy or vice versa.

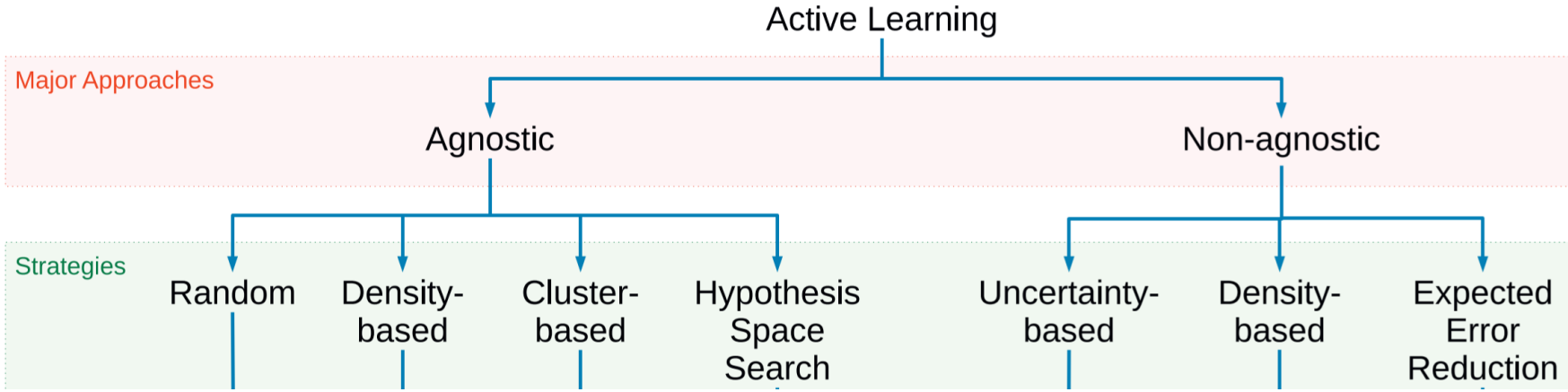
Empirical Investigation of Active Learning Strategies

Davi Pereira-Santos[a,*], Ricardo Bastos Cavalcante Prudêncio[b], André C.P.L.F. de Carvalho[a]

[a] *Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Trabalhador São-carlense Av. 400, São Carlos, São Paulo 13560-970, Brazil*
[b] *Centro de Informática, Universidade Federal de Pernambuco, Jornalista Aníbal Fernandes Av., Recife, Pernambuco 50740-560, Brazil*

Settings

- **Methods**



- **Metrics**

Kappa coefficient $$\kappa = \frac{a - \frac{v_{pred} \cdot v_{exp}}{n}}{1 - \frac{v_{pred} \cdot v_{exp}}{n}}$$

- **Models**

(1) 5NN, (2) C4.5, (3) NB, (4) SVM with RBF kernel, (5) RF

- **Datasets**

75 UCI datasets

Active learning strategies, respective methods and their main aspects.

Approach	Strategy	Method	Search type	Learner
Agnostic	Random sampling	Rnd[a]	exploratory-random	no learner
	Density-based	ATU[b]	exploratory	no learner
	Cluster-based	HS[c]	balanced exploratory/prospective	no learner
	Hypothesis Space Search	SGmulti[d]	delimited exploratory-random	present
Non-agnostic	Uncertainty-based	Mar[a]	prospective	present
	Density-based	TU[e]	balanced exploratory/prospective	present
		HTU[b]	switching exploratory/prospective	present
	Expected Error Reduction	ERE[f], OER[g]	prospective	present

Mean ranking position. *The learner name is in blue bold face and red italicized in its first and last positions, respectively.*

Rank. Pos.	Method	Learner	Rank. Pos.	Method	Learner	Rank. Pos.	Method	Learner
13.36	Mar	**RF**	37.64	ATUman	5NN	42.80	ATUman	NB
15.96	SGmulti	RF	37.85	.HTUmah	5NN	43.20	ATUeuc	NB
16.54	ERE	RF	38.07	ATUmah	5NN	43.65	HTUmah	C4.5
17.39	OER	RF	38.14	HTUeuc	5NN	43.67	ATUmah	NB
22.91	HS	RF	38.82	Mar	SVM	43.72	Rnd	5NN
24.54	TUman	RF	38.83	ERE	5NN	43.72	SGmulti	5NN
25.43	HTUman	RF	38.97	TUmah	5NN	43.78	TUman	C4.5
25.51	TUmah	RF	39.32	ATUman	SVM	43.82	SGmulti	SVM
25.67	TUeuc	RF	39.46	ATUeuc	SVM	43.92	Rnd	*SVM*
25.76	HTUmah	RF	39.49	ERE	**NB**	44.01	ATUman	C4.5
25.83	HTUeuc	RF	39.77	TUman	5NN	44.06	HTUeuc	C4.5
25.83	Rnd	RF	39.94	ATUmah	SVM	44.27	Mar	5NN
29.89	ATUman	RF	40.58	HS	SVM	44.77	TUmah	C4.5
31.42	ATUeuc	RF	40.71	SGmulti	NB	44.91	OER	NB
32.17	ATUmah	*RF*	40.83	HTUmah	NB	45.01	ATUeuc	C4.5
32.93	ERE	**SVM**	41.25	TUeuc	5NN	45.19	TUeuc	C4.5
33.80	HTUeuc	SVM	41.26	HTUeuc	NB	45.45	HS	NB
34.85	HTUmah	SVM	41.35	HTUman	NB	45.79	ATUmah	C4.5
35.12	HTUman	SVM	41.41	TUmah	NB	46.37	OER	C4.5
35.67	TUman	SVM	41.55	TUman	NB	46.38	Rnd	*NB*
36.18	TUmah	SVM	41.70	ERE	**C4.5**	46.62	HS	C4.5
36.57	TUeuc	SVM	42.56	TUeuc	NB	47.43	Mar	C4.5
36.70	OER	SVM	42.64	HS	5NN	47.85	SGmulti	C4.5
37.50	ATUeuc	**5NN**	42.67	HTUman	C4.5	47.94	OER	*5NN*
37.60	HTUman	5NN	42.75	Mar	NB	48.97	Rnd	*C4.5*

- Conclusion

1. Active learning is effective
2. The learning algorithm can be decisive to the overall predictive performance
3. Uncertainty sampling and EER perform well in general

Practical Obstacles to Deploying Active Learning

David Lowell
Northeastern University
lowell.d@husky.neu.edu

Zachary C. Lipton
Carnegie Mellon University
zlipton@cmu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Motivation

- What will happen if we use a **model which is different with the target model to select instance**?

Conduct experiments using different models for active selection & model training

Settings

- **Tasks**

    1. Text classification
    2. Named Entity Recognition

- **Methods**

    1. Uncertainty Sampling
    2. QBC
    3. (Bayesian AL by Disagreement) BALD

- **Metrics**

    F1/accuracy

Dataset	# Classes	# Documents	Examples per Class
Movie Reviews	2	10662	5331, 5331
Subjectivity	2	10000	5000, 5000
TREC	6	5952	1300, 916, 95, 1288, 1344, 1009
Customer Reviews	2	3775	1368, 2407

Table 3: Text classification dataset statistics.

Text classification

Successor	pool %	i.i.d.	Acquisition model: Uncertainty SVM	Uncertainty CNN	Uncertainty LSTM	QBC SVM	QBC CNN	QBC LSTM	BALD CNN	BALD LSTM
Movie reviews										
SVM	10	65.3	**65.3**	65.8	65.7	**64.9**	64.9	65.1	64.9	65.2
	20	68.2	**69.0**	69.4	68.9	**68.1**	68.4	68.7	68.5	69.0
CNN	10	65.0	65.3	**65.5**	65.4	64.8	**65.1**	64.7	**65.1**	64.9
	20	69.4	69.1	**69.5**	69.5	68.5	**69.1**	69.1	**68.3**	69.1
LSTM	10	63.0	62.0	62.5	**63.1**	61.9	61.9	**62.6**	61.7	**62.2**
	20	67.2	65.1	65.8	**67.0**	65.4	65.7	**66.8**	65.6	**67.1**
Subjectivity										
SVM	10	85.2	**85.6**	85.3	85.5	**85.4**	85.0	85.4	85.8	85.4
	20	87.5	**87.6**	87.4	87.6	**87.7**	87.0	87.5	87.0	87.6
CNN	10	85.3	85.2	**86.3**	86.0	85.3	**86.0**	85.7	**86.2**	85.7
	20	87.9	87.6	**88.4**	88.6	88.4	**88.5**	88.6	**88.6**	88.3
LSTM	10	82.9	82.7	82.7	**84.1**	83.3	83.7	**84.8**	83.1	**84.2**
	20	86.7	86.3	85.8	**87.6**	86.9	87.0	**87.7**	84.7	**87.0**

Models trained on foreign actively acquired datasets tend to **underperform** those trained on i.i.d. datasets

Text classification

Successor	pool %	i.i.d.	Acquisition model: Uncertainty SVM	Uncertainty CNN	Uncertainty LSTM	QBC SVM	QBC CNN	QBC LSTM	BALD CNN	BALD LSTM
TREC										
SVM	10	68.5	**68.3**	66.8	68.5	**68.1**	63.1	64.9	68.2	68.3
	20	74.1	**74.7**	73.2	74.3	**73.7**	71.6	71.2	74.1	74.1
CNN	10	70.9	70.5	**69.0**	70.0	67.4	**62.8**	69.5	**71.0**	70.5
	20	76.1	77.7	**77.3**	78.0	76.5	**73.7**	76.3	**79.8**	77.7
LSTM	10	65.2	64.5	63.6	**63.8**	61.7	60.1	**64.6**	64.1	**64.5**
	20	71.5	72.7	71.0	**73.3**	71.4	69.9	**71.8**	72.9	**72.6**
Customer reviews										
SVM	10	68.8	**70.5**	70.3	68.5	**70.5**	69.5	64.6	70.0	69.2
	20	73.6	**74.2**	72.9	71.1	**73.8**	72.6	65.7	73.5	71.7
CNN	10	70.6	70.9	**71.7**	68.2	71.5	**71.4**	63.8	**72.2**	68.4
	20	74.1	74.5	**74.8**	71.5	74.9	**74.9**	65.2	**75.3**	71.3
LSTM	10	66.1	67.2	65.1	**65.9**	65.0	64.8	**64.0**	65.2	**65.4**
	20	68.0	66.6	66.5	**66.3**	66.3	66.4	**65.4**	68.3	**68.0**

Only **37.5%** of the tabulated data points representing dataset transfer (in which acquisition and successor models differ) **outperform** the i.i.d. baseline.

Named Entity Recognition

Successor	pool %	i.i.d.	Acquisition Model: Uncertainty CRF	Uncertainty BiLSTM-CNN	BALD BiLSTM-CNN	QBC CRF	QBC BiLSTM-CNN
				CoNLL			
CRF	10	69.2	**70.5**	70.2	70.3	**70.3**	70.0
	20	73.6	**74.4**	74.0	74.1	**74.5**	74.1
BiLSTM-CNN	10	87.4	87.4	**87.8**	**88.0**	87.5	**87.7**
	20	89.1	89.6	**89.6**	**89.8**	89.2	**89.5**

Successor	pool %	i.i.d.	Acquisition Model: Uncertainty CRF	Uncertainty BiLSTM-CNN	BALD BiLSTM-CNN
				OntoNotes	
CRF	10	73.8	**75.5**	75.4	75.3
	20	77.6	**79.1**	78.7	78.7
BiLSTM-CNN	10	82.6	83.1	**83.1**	**83.2**
	20	84.6	85.2	**84.9**	**85.1**

Table 6: F1 measurements for the NER task, with training sets comprising 10% and 20% of the training pool.

感谢聆听！