Combining Crowd and Expert Labels using Decision Theoretic Active Learning

An T. Nguyen [1] Byron C. Wallace Matthew Lease

University of Texas at Austin

HCOMP, 2015

[1] Presenter

The Problem: Label Collection

- Have some unlabeled data.
- Want labels
- of high quality at low cost.

The Problem: Label Collection

- Have some unlabeled data.
- Want labels
- of high quality at low cost.

Finite Pool Setting

- Care about label quality of current data.
- Dont care (much) about future data.

Some Solutions

Some Solutions

- Hire a domain expert to give labels.

Some Solutions

- Hire a domain expert to give labels.
- Crowdsource the labeling.

Some Solutions

- Hire a domain expert to give labels.
- Crowdsource the labeling.
- Build a Prediction Model (Classifier).

Some Solutions

- Hire a domain expert to give labels.
- Crowdsource the labeling.
- Build a Prediction Model (Classifier).

Our work: A principled way to combine these:

Some Solutions

- Hire a domain expert to give labels.
- Crowdsource the labeling.
- Build a Prediction Model (Classifier).

Our work: A principled way to combine these:

- Which item ? Which labeler?
- How to use classifier ?

Method: Previous work

Roy and McCallum 2001

- 'Optimal' Active Learning.

Method: Previous work

Roy and McCallum 2001

- 'Optimal' Active Learning.
- Select item to get label by

Method: Previous work

Roy and McCallum 2001

- 'Optimal' Active Learning.
- Select item to get label by

    1. Consider each item
    2. Consider each possible label.

Method: Previous work

Roy and McCallum 2001

- 'Optimal' Active Learning.
- Select item to get label by

    1. Consider each item
    2. Consider each possible label.
    3. Add that (item, label) to the training set
    4. Retrain and Evaluate.

Method: Previous work

Roy and McCallum 2001

- 'Optimal' Active Learning.
- Select item to get label by

    1. Consider each item
    2. Consider each possible label.
    3. Add that (item, label) to the training set
    4. Retrain and Evaluate.
    5. Weight outcomes by (predictive) probabilities
    6. Select one with best expected outcome.

Method: Previous work

Roy and McCallum 2001

- 'Optimal' Active Learning.
- Select item to get label by

    1. Consider each item
    2. Consider each possible label.
    3. Add that (item, label) to the training set
    4. Retrain and Evaluate.
    5. Weight outcomes by (predictive) probabilities
    6. Select one with best expected outcome.

- Basically one-step look-ahead
- A (perhaps) better name: Decision Theoretic Active Learning.

Method: Our ideas

The key idea: Extend their algorithm to include expert/crowd/classifier.

Method: Our ideas

The key idea: Extend their algorithm to include expert/crowd/classifier.

- Consider (item, label, **labeler**).

Method: Our ideas

The key idea: Extend their algorithm to include expert/crowd/classifier.

- Consider (item, label, **labeler**).
- Have a Crowd Accuracy Model: $Pr(\text{True L}|\text{Crowd L}) = ?$

Method: Our ideas

The key idea: Extend their algorithm to include expert/crowd/classifier.

- Consider (item, label, **labeler**).
- Have a Crowd Accuracy Model: $Pr(\text{True L}|\text{Crowd L}) = ?$

Strategy: Loss Prediction/Minimizaion

- Loss for expert labels $= 0$
- Predict Loss for crowd labels
- Predict Loss for classifier's prediction

Method: Our ideas

The key idea: Extend their algorithm to include expert/crowd/classifier.

- Consider (item, label, **labeler**).
- Have a Crowd Accuracy Model: $Pr(\text{True L}|\text{Crowd L}) = ?$

Strategy: Loss Prediction/Minimizaion

- Loss for expert labels $= 0$
- Predict Loss for crowd labels
- Predict Loss for classifier's prediction
- Predict Loss Reduction after adding a label by a labeler.

Decision Criteria: Loss Reduction/Cost

Evaluation: Application

Evidence Based Medicine (EBM)

aims to inform patient care using the entirety of the evidence.

Evaluation: Application

Evidence Based Medicine (EBM)

aims to inform patient care using the entirety of the evidence.

Biomedical Citation Screening

is the first step in EBM: identify relevant citations (paper abstracts, titles, keywords ...).

Evaluation: Application

Evidence Based Medicine (EBM)

aims to inform patient care using the entirety of the evidence.

Biomedical Citation Screening

is the first step in EBM: identify relevant citations (paper abstracts, titles, keywords ...).

Two characteristics:

- Very imbalanced (2-15% positive).
- Recall a lot more important than Precision.

Evaluation: Application

Evidence Based Medicine (EBM)

aims to inform patient care using the entirety of the evidence.

Biomedical Citation Screening

is the first step in EBM: identify relevant citations (paper abstracts, titles, keywords ...).

Two characteristics:

- Very imbalanced (2-15% positive).
- Recall a lot more important than Precision.

The expert

- MD, specialist
- very expensive, paid 100 times a crowdworker.

Evaluation: Data

Four Biomedical Citation Screening Datasets

Evaluation: Data

Four Biomedical Citation Screening Datasets

- Have expert gold labels.
- Have crowd labels (5 for each item) ...
- collected via Amazon Mechanical Turk.

Evaluation: Data

Four Biomedical Citation Screening Datasets

- Have expert gold labels.
- Have crowd labels (5 for each item) ...
- collected via Amazon Mechanical Turk.

Strategy to use

1. Test/Refine our methods using **only** the First & Second.

Evaluation: Data

Four Biomedical Citation Screening Datasets

- Have expert gold labels.
- Have crowd labels (5 for each item) ...
- collected via Amazon Mechanical Turk.

Strategy to use

1. Test/Refine our methods using **only** the First & Second.
2. **Finalize** all details (e.g. hyper-parameters).

Evaluation: Data

Four Biomedical Citation Screening Datasets

- Have expert gold labels.
- Have crowd labels (5 for each item) ...
- collected via Amazon Mechanical Turk.

Strategy to use

1. Test/Refine our methods using **only** the First & Second.
2. **Finalize** all details (e.g. hyper-parameters).
3. Test on the Third & Forth.

Evaluation: Data

Four Biomedical Citation Screening Datasets

- Have expert gold labels.
- Have crowd labels (5 for each item) ...
- collected via Amazon Mechanical Turk.

Strategy to use

1. Test/Refine our methods using **only** the First & Second.
2. **Finalize** all details (e.g. hyper-parameters).
3. Test on the Third & Forth.
4. Purpose: See how it performs on **real future data**.

Evaluation: Setup

Active Learning Baseline: Uncertainty Sampling (US)

Select item with probability closest to 0.5

Evaluation: Setup

Active Learning Baseline: Uncertainty Sampling (US)

Select item with probability closest to 0.5

Compare Four Algorithms

- US-Crowd: use only crowd labels.
- US-Expert: use only experts.
- US-Crowd+Expert: Crowd first. Expert if disagree.
- Decision Theory: our method.

Evaluation: Metric

Compare collected labels vs. gold labels

Evaluation: Metric

Compare collected labels vs. gold labels

Collected labels includes:

- Expert labels.
- Crowd (Majority Voting)
- Classifier predictions (trained on crowd & expert labels)

Evaluation: Metric

Compare collected labels vs. gold labels

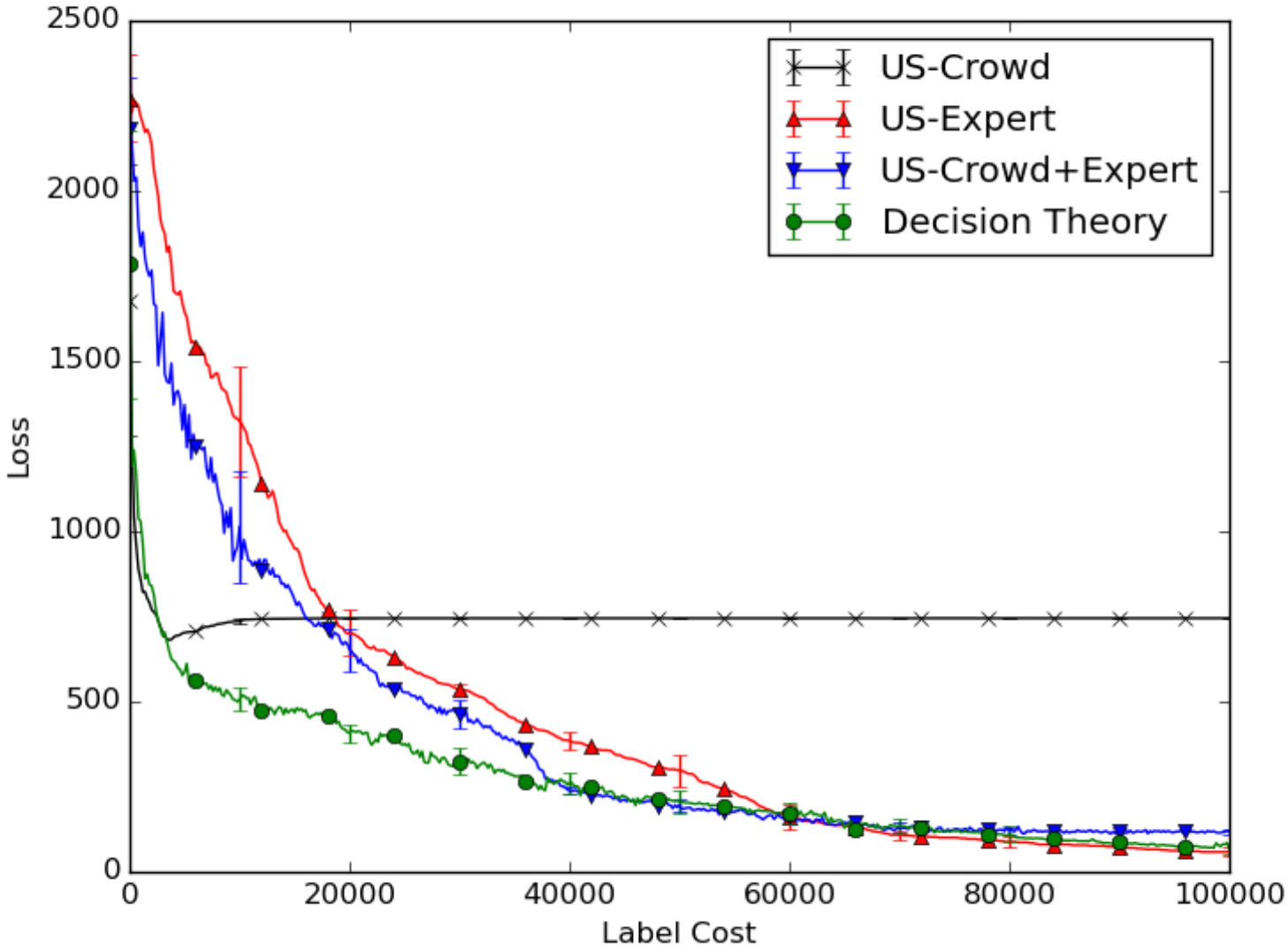
Collected labels includes:

- Expert labels.
- Crowd (Majority Voting)
- Classifier predictions (trained on crowd & expert labels)

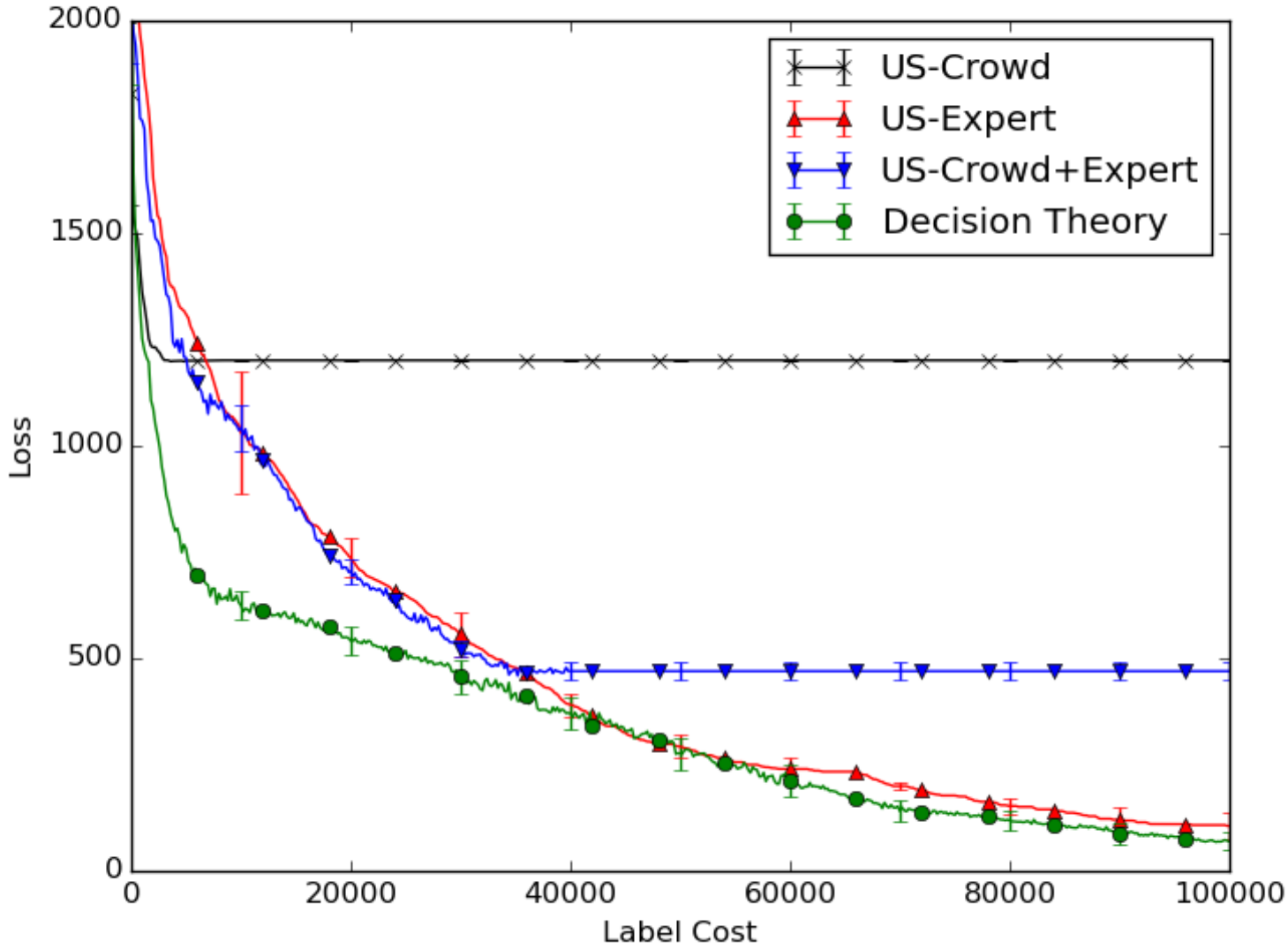
We present: Cost-Loss Learning Curve

- One Expert Label $= 100$, One Crowd Label $= 1$.
- Loss $=$ # False Positive $+ 10$ # False Negative.

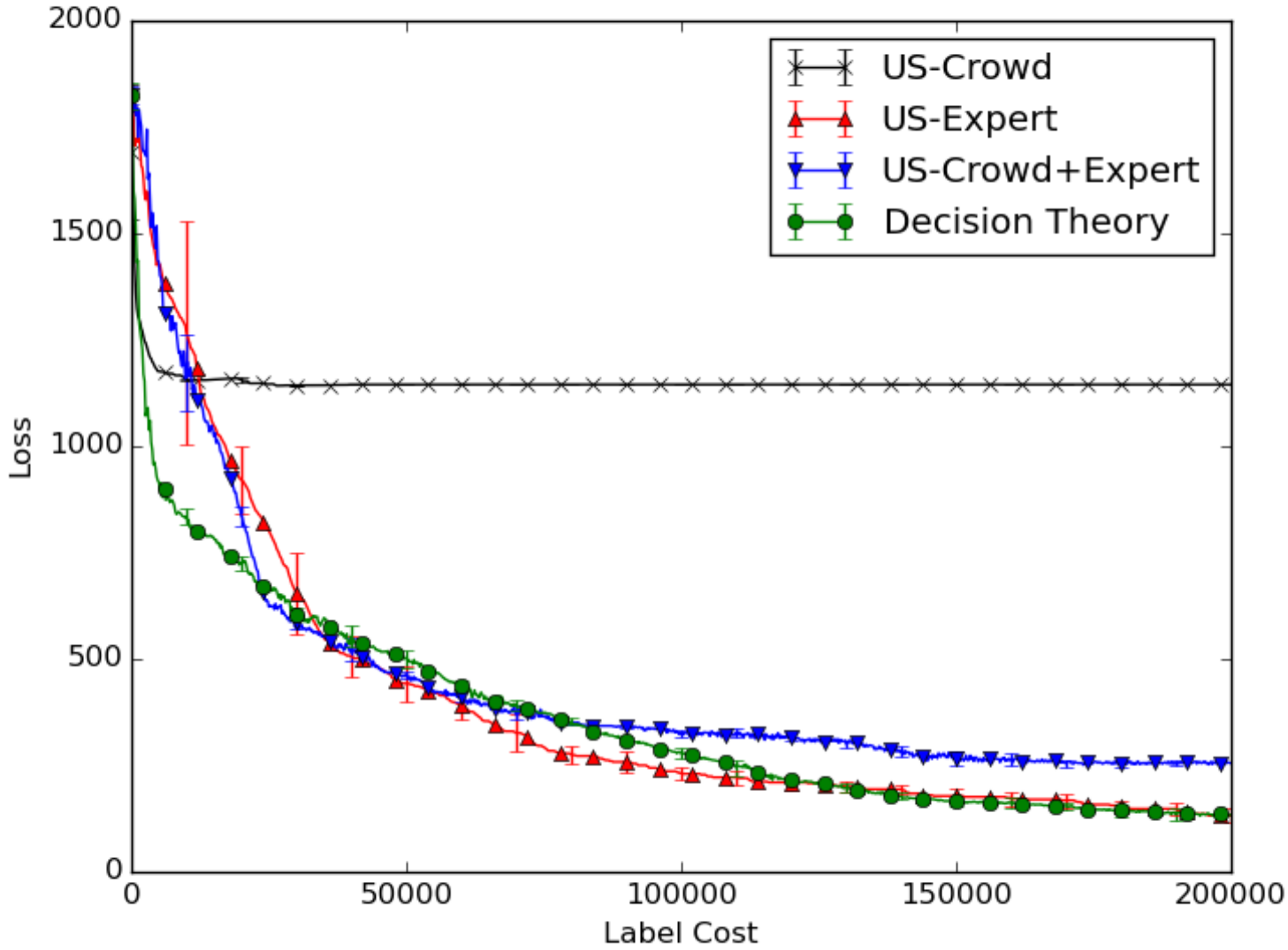
Evaluation: Result: First Dataset



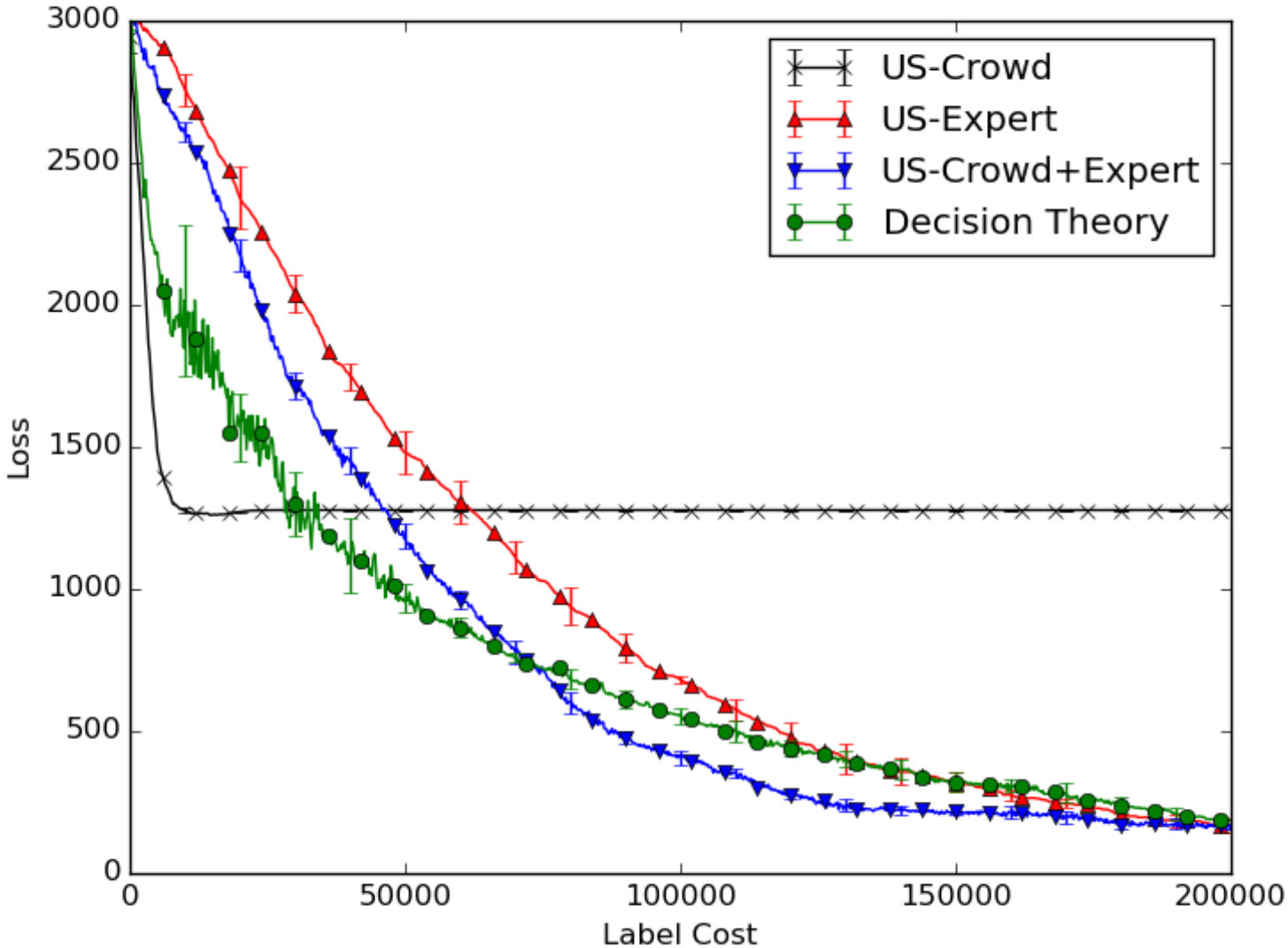
Evaluation: Result: Second Dataset



Evaluation: Result: Third (real future) Dataset



Evaluation: Result: Forth (real future) Dataset



Discussion

Our method

- Overall effective. Consistenly good in the beginning.
- On ‘real future datasets’: lose slightly at some points.

Discussion

Our method

- Overall effective. Consistenly good in the beginning.
- On 'real future datasets': lose slightly at some points.

Future work

- Better worker model.
- Multi-step lookahead.
- Quality Assurance/Guarantee.

Summary

We have presented

- High level ideas of our method.
- Evaluation and Results

Summary

We have presented

- High level ideas of our method.
- Evaluation and Results

We have omitted

- Full algorithms. Impplementation details.
- Heuristics to make this fast.
- Crowd Model. Active Sampling Correction.
- More results.

Summary

We have presented

- High level ideas of our method.
- Evaluation and Results

We have omitted

- Full algorithms. Impplementation details.
- Heuristics to make this fast.
- Crowd Model. Active Sampling Correction.
- More results.
- See the paper.

Summary

We have presented

- High level ideas of our method.
- Evaluation and Results

We have omitted

- Full algorithms. Impplementation details.
- Heuristics to make this fast.
- Crowd Model. Active Sampling Correction.
- More results.
- See the paper.

Question?

References I

Roy, Nicholas and Andrew McCallum (2001). "Toward Optimal Active Learning through Sampling Estimation of Error Reduction". In: *In Proc. 18th International Conf. on Machine Learning*.