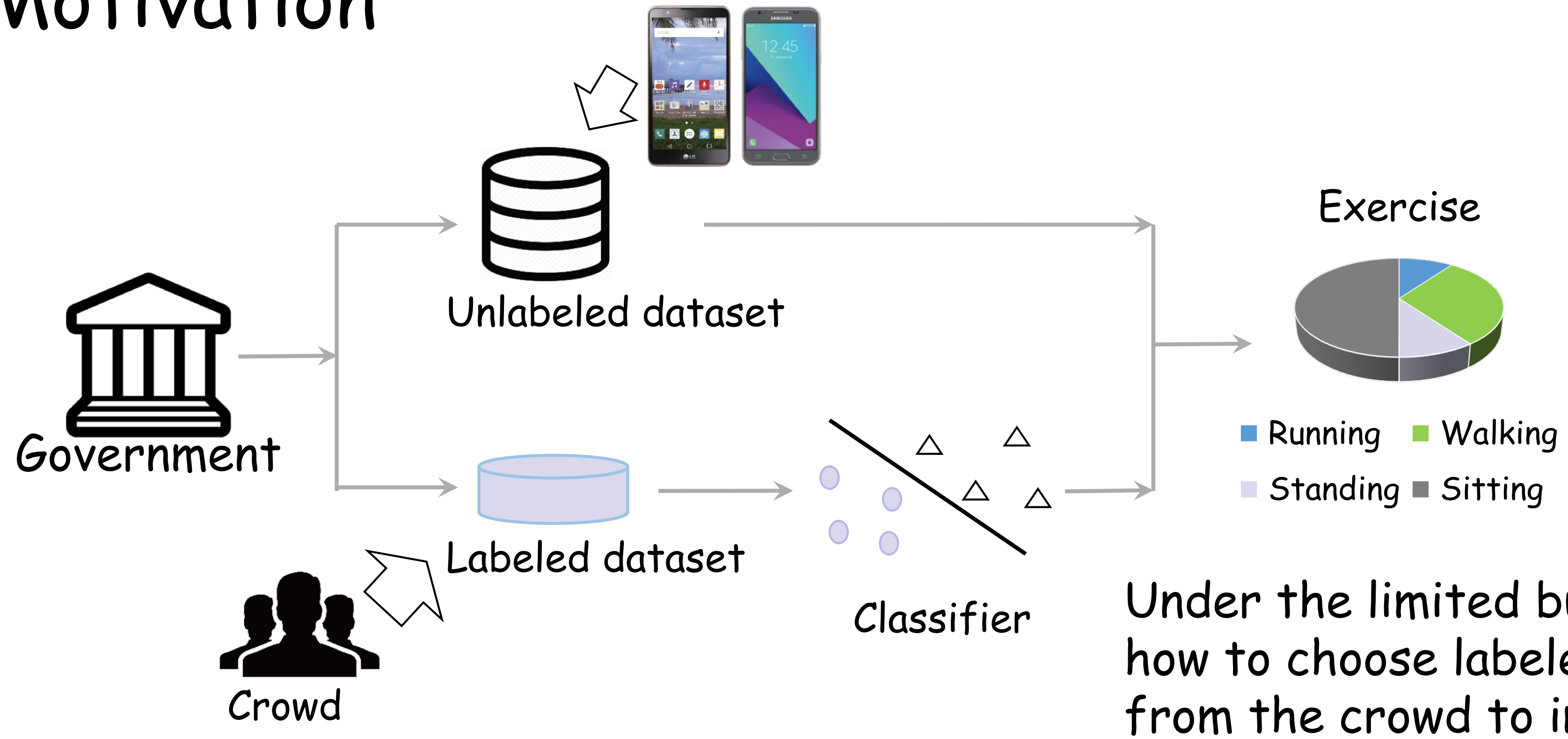
Data Utility Maximization When Leveraging Crowdsensing in Machine Learning

Juan Li, Jie Wu, and Yanmin Zhu

Shanghai Jiao Tong University

Temple University

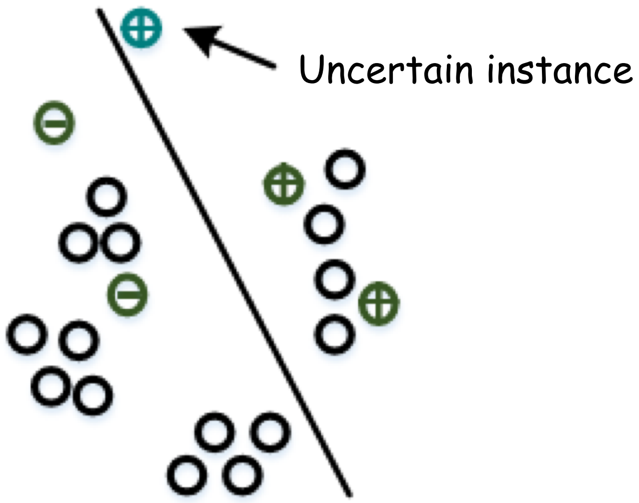
Motivation



Under the limited budget, how to choose labeled data from the crowd to improve the accuracy of the classifier most?

Uncertainty

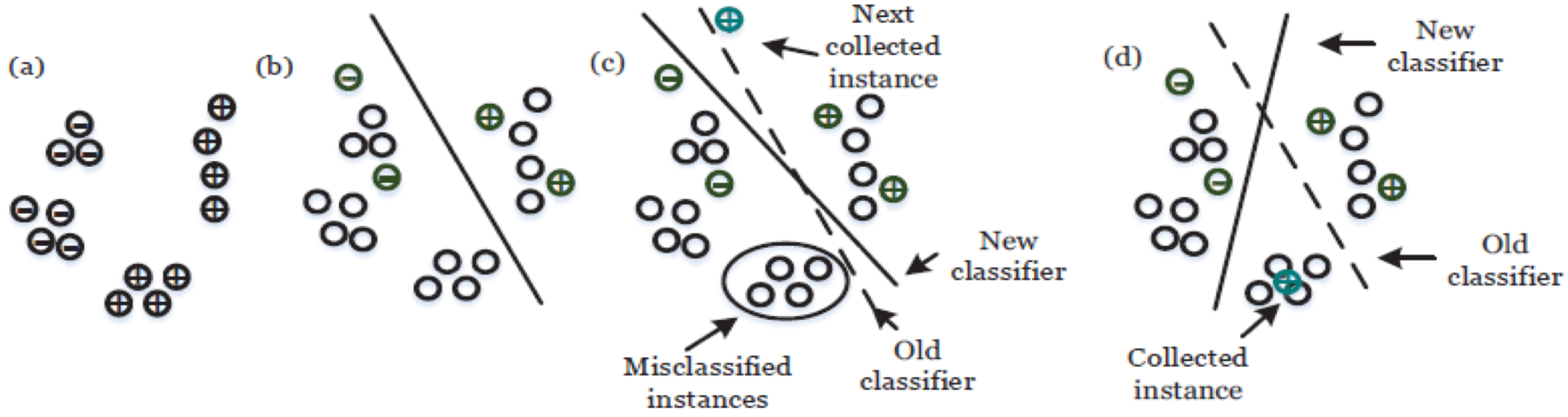
Confidence-based, margin-based and entropy-based uncertainty measures



Margin-based measure

Label y_1 and y_2 are the first and second most likely predictions for instance x under the classification model $f(\Theta)$.
The margin is $m = P(y_1|x,\Theta) - P(y_2|x,\Theta)$.
The uncertainty of the model about x is $u(\mathrm{x}) = 1 - m$.

Weighted Density



The unlabeled data set Q and true labels which are actually unknown.

The current training set and the current classifier.

Collecting the most uncertain data instance.

Collecting the instance with the highest weighted density.

Marginal effect $V(x) = u(\mathrm{x})\times\sum_{x'\in Q} u(x')sim(x,x')$

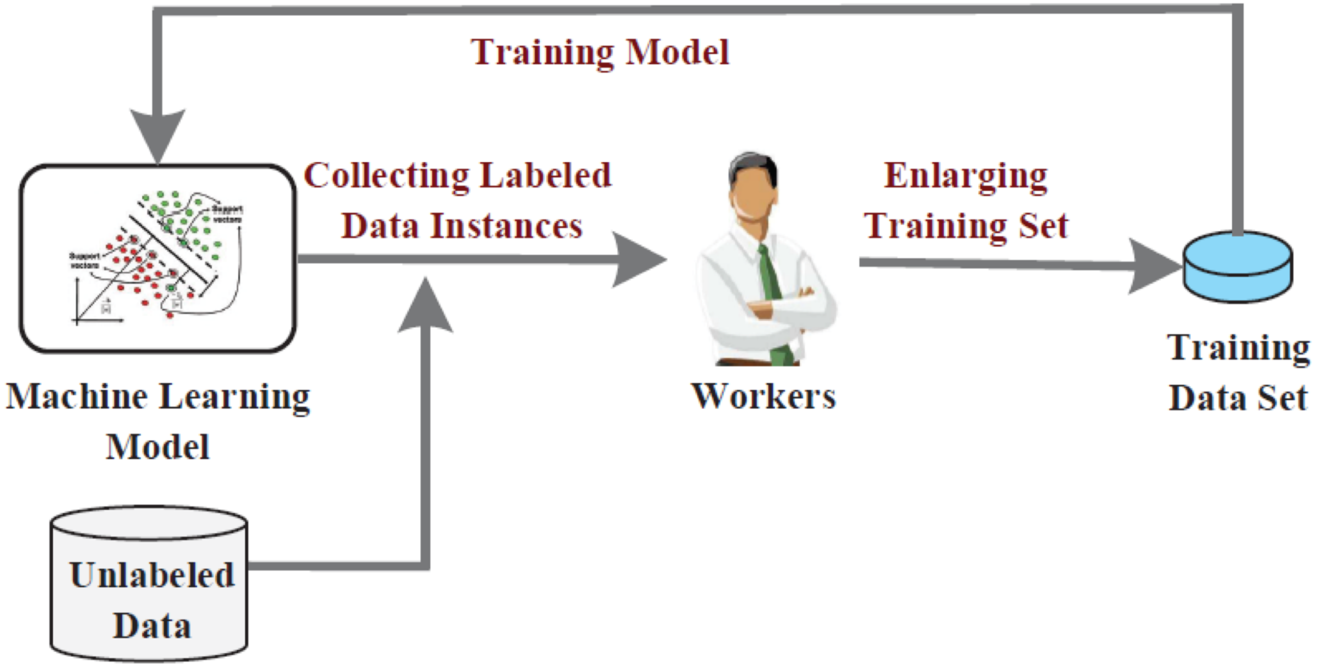
Crowdsensing Framework & Problem

In each round, we try to maximize data utility under the budget of a round.

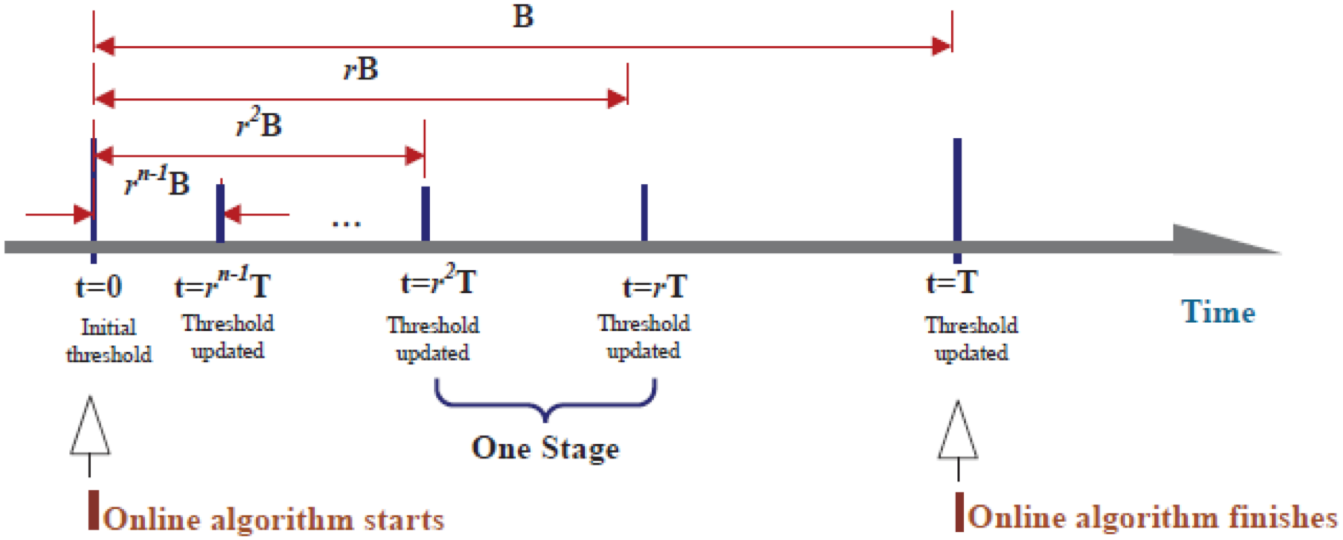
$$max\, V(S)$$

$$s.t. \sum_{x_i \in S} c_i \leq B$$

$$s.t. V(S) = \sum_{x_i \in S} V(x)$$



Online Algorithm



$r \in (0,1)$

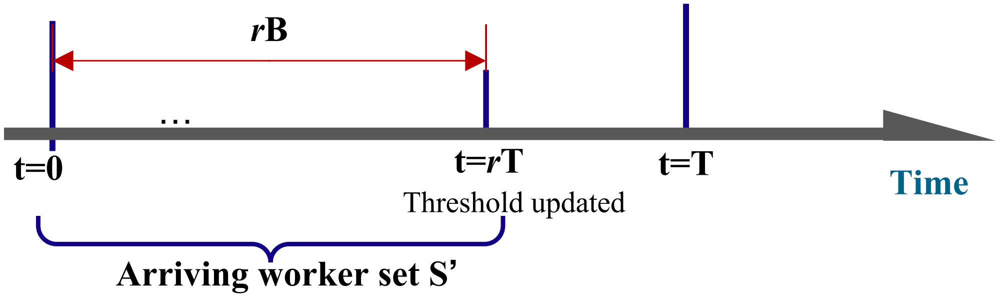
Marginal contribution: $V_i(S) = V(S \cup x_i) - V(S)$
Marginal efficiency: $V_i(S)/c_i$

In each stage, we recruit the coming worker if
1)the marginal efficiency is not less than the threshold.
2)the budget in that stage is not run out of.

We update the threshold at the end of each stage.

Threshold Updating

Learn from the past



We choose an optimal worker set $W \in S'$ to maximize data utility. The efficiency is $\mathrm{e} = V(W)/(rB)$. The threshold is e/δ.

We continuously choose the instance with the largest marginal efficiency until the budget is run out of. We use $V(W' \cup \{x\})/(1-1/e)$ as the estimation of the optimal data utility.

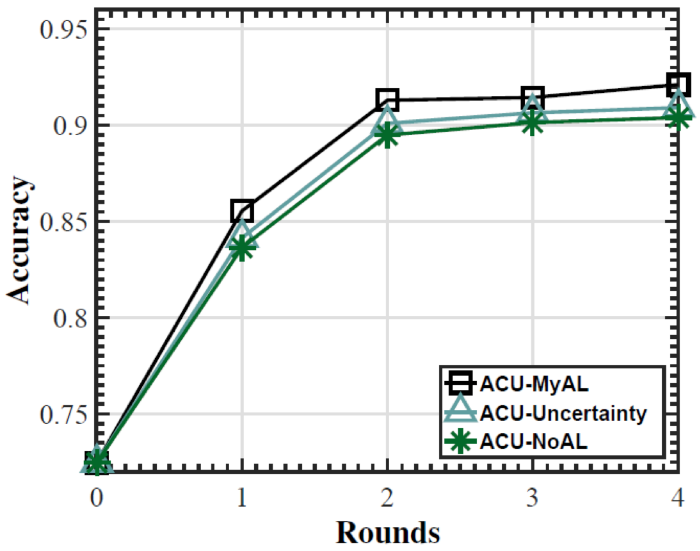
The competitive ratio is 0.1218 if

1) we set $\delta = 4.0648$ and $r = 0.4390$;
2) the contribution of one instance is infinitely small compared with the total data utility achieved by our algorithm;
3) workers arrive randomly.

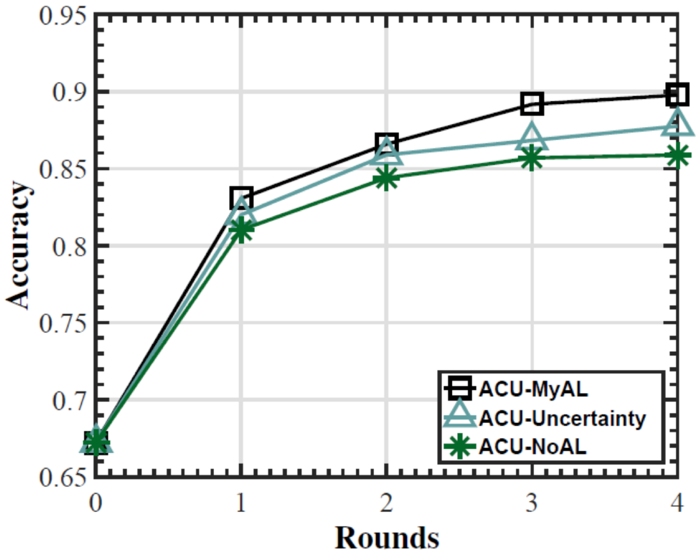
Evaluation

Accuracy achieved in each round under different data utility models

(Human Activity Recognition Using Smartphones Dataset)



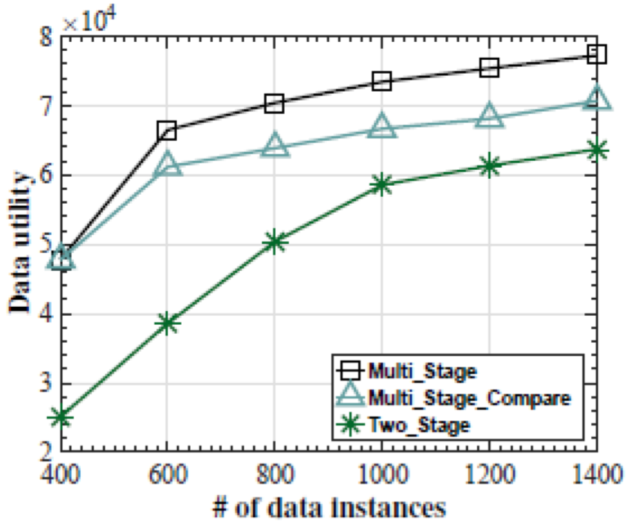
Two-class classification(logistic regression)



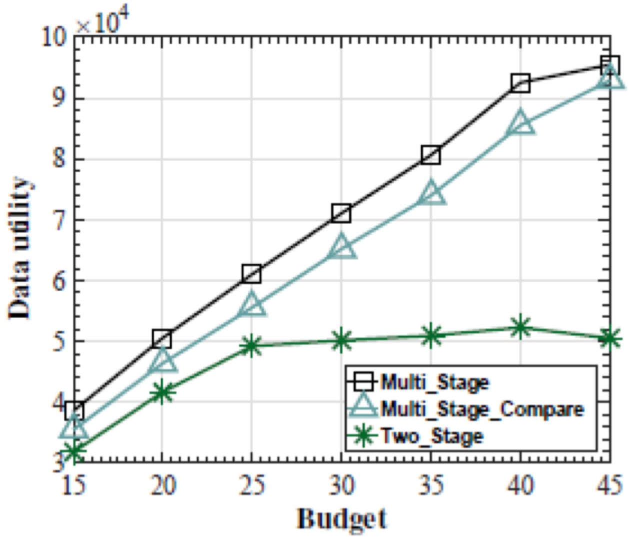
Multiclass classification(SVM)

Motivation → Data Utility → Framework & Problem → Algorithm → Evaluation

Evaluation



Data utility vs. # of coming instances under different algorithms



Data utilities vs. budget under different online algorithms.

Conclusion

1) In this paper, we have studied the data utility maximization problem under the budget constraint when leveraging crowdsensing in machine learning.

2) We come up with a novel data utility model to bridge the gap between the performance of the trained model and the collected instances.

3)We further design a fair online algorithm and achieve a non-trivial competitive ratio.