SIGCHI Extended Abstracts Sample   
File: Note Initial Caps

*Please do not modify this text block until you receive explicit instructions.*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [Permissions@acm.org](mailto:Permissions@acm.org).

*CONF '22,* Jan 1 - Dec 31 2022, Authorberg.

Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-xxxx-yyyy-z/zz/zz…$zz.00.

unique doi string will go here

# Abstract

Masahiro Honjo

Tokyo University,

Hongo, Bunkyo-ku, Japan

honjo@gmail.com

Second Author

VP, Authoring

Authorship Holdings, Ltd.

Awdur SA22 8PP, UK author2@author.ac.u

Crowdsourcing is well-studied as one way of gathering training data for machine learning. Furthermore, crowdsourcing with active learning can be gather efficiency at the point of cost. In this paper, we suggest a model of the crowdsourcing with the active learning that we assumed that is applicable to a real platform, which is combination with labeler which is very higher in reliability and cost than crowd workers, and using conviction of workers for learning.

# Author Keywords

Crowdsourcing Active learning

# ACM Classification Keywords

まだ

# Introduction

Data analysys or classification is well-studied with machine learning. In many cases of machine learning, we face to big problem that how to gather much training data, which are set of question s and answers, because if we leave making all data to a few experts who have high skill, it is very costly points of manetary and temporally cost.

Crowdsourcing can solve this problem to some extent ,which we can reduce their cost to request mass corwd workers much tasks. However, crowdsourcing has problem that its quality of works are very capricious with along ability of crowd workers. So crowdsourcing are well-studied many points of view to enhance accuracy. A method of these study, we noticed a method of crowdsourcing combined with avtive learning to gather data efficiency.

In this paper we suggest method of model of crowdsourcing combined with active learning to balance with accuracy,manetary cost, and temporally cost ,which is considered that we can apply to real platform like Amazon Mechanical Turk.

This paper is structured as follows, in next section we discuss related works. Next, we present our method and compare with related works. At the end, we suggest experimental setup and discuss a future work.

# Related Works

## Active Learning

Active learning is well-studied as one algorithm of machine learning(). Machine learning need much training data, however, it is well known that accurate annotation, which classify each datum manually to make training data take long time(). To solve this problem, in active learning, learner choose a datum that should be learn next interigently to enhance classifier accuracy. To achieve this, many strategy for choosing data were suggested by now.

Active learning can be classified by way of gathering data, and choosing data according to ().

First, active learning can be classified 3 types by way of gathering data.

* Membership query synthesis
* Stream-based selective sampling
* Pool-based sampling

In “Membership query synthesis'', learner makes data that it wants to ask by itself. In “stream-based selective sampling'', data streams and learner judges if it should ask. In “pool-based sampling'', learner choose a datum that should be asked in un-labeled data finite pool. In this study, we deal with only “pool-based sampling''

Second, active learning can be classified 6 types by way of choosing data.

* uncertainty sampling
* Query-By-Committee
* Expected Model Change
* Expected Error Reduction
* Variance Reduction
* Density-Weighted Methods

We discuss ``unsertainty sampling'' and ``expected error reduction'' as follow

## uncertainty Sampling.

Perhaps,the most major strategy of 6 strategy is ``Unceratinly Sampling'', which is asked most ``uncertain'' data. In ``uncertainty sampling'', we assamped using probabilistic classification model, and we can define ``uncertainty'' variously. For example, we choose least confident data ,which has minimum probability of best label. While, in margin sampling, we choose maximum margin data ,which is difference between best and second-best label is maximum. In ``entropy-based approach'', we choose data whose distribution's entropy is maximum. we should choose best ``uncertainty'' case-by-case.

Although ``uncertainty sampling'' is very simple strategy, it is very powerful strategy.

In this paper,so we discuss and compare with proposal method as baseline.

## Expected Error Reduction

``Expected error reduction'' is like the algorithm as follow:

1. Train classifier using current training data, and estimating probablity for each data.
2. Estimate current Loss.
3. For each un-labeled data, it assamped that it was labeled each label.
4. Retrain classifier, and expected Loss.
5. Choose data which have maximum expectation of loss reduction, and ask.
6. Back to [1], repeat.

``Expected error reduction'' require tremendous computational cost in space, because we have to retarain classifier {(number of data)$\times$(number of class)} simply.

In step [3], we estimate loss as follow formula:

- \sum\_{x \in U} \sum\_{y} P(y|x,T \cup \{(x\_{i},y\_{i})\}) \log P(y|x,T \cup \{(x\_{i},y\_{i})\})

($T \cup \{(x\_{i},y\_{i})\})$ means that we calculate on traning data added a data, $\{(x\_{i},y\_{i})\}$, to current one.

we choose $x\_{i}$ from unlabeled data,$R(x\_{i},y\_{i};L)$ is defined formula (\ref{equ:eer})

argmin(E\_{y\_{i}}[R(x\_{i},y\_{i};L) = \sum\_{y\_{i}} P(y\_{i}|x\_{i};L)R(x\_{i},y\_{i};L)

``Expected error redution'' is useful when we have a number of action type. For example, asking crowd worker or expert. However, computational cost in space is to be directly propotional to number of action type.

An and Wallace were combined crowdsourcing and ``expected error reduction'' , which they called ``optimal active learning'' to judge if asking to crowd-worker or expert. we introduce their method as follow.

## Combining Crowd and Expert Labels Using Decision Theoretic Active Learning

An considered combined with crowd workers and expert by using ``excepted error reduction'' .

In their method, they calculate expected error reduction per each cost, crowd workers and expert, and choose data and action, follow as formula (\ref{equ:an}) on behalf of formula (\ref{equ:eer}).

\cfrac{\sum\_{x \in U} \sum\_{y} P(y,w|x,T \cup \{(x\_{i},y\_{i})\}) Loss(y|x,T \cup \{(x\_{i},y\_{i})\})}{cost(w)}

$w$ means worker type.

However, it is very tremendous computational cost in space to execute fheir method as it is, as mentioned above.

So, they use various approximation as follow

* when expected error by formula \ref{equ:eer}, using a part of $({x\_i},{y\_i})$
* expert can label only a data that crowd worker labeled.

Their simulation showed that their method superiored to baseline, uncertainty sampling, when total budget is limited total budget in their paper.

However,in their paper,they got data which crowd workers labeled in advance, and their algiorithm ask crowd-workers 1 data at 1 time.

In our suggestion, we extend their method to fit better in real platform of crowdsourcing.