K Vipul Arya
26 July 2023 17:15

- My review of the paper !!
- Paper - https://www.lamda.nju.edu.cn/publication/ijcai17ceal.pdf
- Active learning is a machine learning technique that involves selecting the most informative examples to be labeled by an oracle (a human annotator or a pre-labeled dataset) and then using these labeled examples to train a classification model.
- Active learning aims to minimize the number of labeled examples needed to achieve a certain level of accuracy in the classification model.
- In traditional active learning, the oracle is assumed to be a single, high-quality annotator with a fixed cost per label. However, in real-world scenarios, the quality and cost of labeling can vary widely depending on the annotator and the task.
- For example, male customers may be more confident in rating a shaver than a skirt, and some annotators may be more accurate or faster than others.
- The authors propose a novel criterion for active selection on both instances and labelers to address these challenges. The criterion evaluates the cost-effectiveness of instance-labeler pairs, ensuring that the selected instance helps improve the classification model and the selected labeler can assign an accurate label for the instance with a low cost.
- The authors conduct experiments on several real crowdsourcing datasets to compare the proposed approach with other active learning methods.
- The results show that the proposed approach outperforms other methods in terms of accuracy and cost-effectiveness.
- In summary, this paper proposes a new approach to active learning that takes into account the diversity of labelers and their costs and demonstrates its effectiveness through experiments on real-world datasets.

ABSTRACT

- The abstract begins by highlighting the challenge of label acquisition for unlabeled data, which is usually expensive because it requires the participation of human experts.
- The paper aims to reduce the labeling cost by selecting only the most valuable data for their class assignments, using a technique called active learning.
- The authors propose a novel active selection criterion that takes into account the diversity of labelers and their costs.
- This criterion ensures that the selected instance helps improve the classification model and the selected labeler can assign an accurate label for the instance with a low cost.
- The paper presents experiments on both UCI and real crowdsourcing datasets to demonstrate the superiority of the proposed approach in selecting cost-effective queries.
- The results show that the proposed approach outperforms other active learning methods in terms of accuracy and cost-effectiveness.

INTRODUCTION

- The introduction begins by highlighting the challenge of label acquisition for unlabeled data, which is usually expensive because it requires the participation of human experts.
- The goal of active learning is to reduce the labeling cost by selecting only the most valuable instances for their class assignments.
- The authors then introduce the concept of diverse labelers, which refers to the fact that different labelers usually have diverse expertise, and thus it is likely that labelers with a low overall quality can provide accurate labels in some specific instances.
- This observation suggests that globally selecting one labeler for all instances may result in unnecessary high costs, and one better choice is to adaptively select the most cost-effective labeler for each instance.
- To address these challenges, the authors propose a novel active selection criterion that takes into account the diversity of labelers and their costs.
- This criterion ensures that the selected instance helps improve the classification model and the selected labeler can assign an accurate label for the instance with a low cost.

RELATED WORK

- The authors begin by noting that active learning has been well-studied in recent years, and many algorithms have been proposed to design different criteria for query selection.
- The goal of these criteria is to select the most informative examples to be labeled by an oracle (pre-labeled dataset) and then use these labeled examples to train a classification model.

The authors then categorize active learning methods into three groups based on the query criteria used:

1. The methods querying most informative instances: These methods select instances that are most uncertain or have the highest

expected error reduction. Examples of such methods include uncertainty sampling and expected error reduction-based sampling.

2. The methods querying most representative instances: These methods select instances that are most representative of the data distribution. Examples of such methods include clustering-based sampling and density-based sampling.

3. The methods that simultaneously consider informativeness and representativeness: These methods aim to balance the trade-off between informativeness and representativeness. Examples of such methods include those based on active learning with clustering and active learning with multiple views.

- However, these approaches have simply assumed an identical cost for all labelers, which may result in an expensive solution.
- Authors emphasize the diversity of labelers on both their expertise and query cost and propose a novel active selection criterion that takes into account the cost-effectiveness of instance-labeler pairs.

THE PROPOSED ALGORITHM

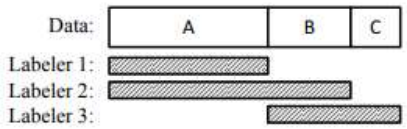

Figure 1: Multiple labelers with diverse expertise. The textured boxes indicate on which parts of data each labeler can give accurate labels.

- At each iteration of active learning, the algorithm selects the most cost-effective pair (x*, a*), where x* is an instance from the unlabeled set U and a* is a labeler from a set of diverse labelers A.
- The cost-effectiveness of a pair is evaluated based on two criteria:

1. The instance x* helps improve the classification model. This criterion is measured by the uncertainty of x* concerning the current model.

2. The labeler a* can provide an accurate label for x* with a relatively low cost. This criterion is measured by the cost of a*.

- Once the most cost-effective pair (x*, a*) is selected, the algorithm queries the label of x* from a* and removes x* from the unlabeled set U. The queried label ˆy* is added to the labeled set L along with x*. The classification model is then retrained on the expanded labeled set L.
- The updated model is evaluated on a hold-out test set, and the active querying and model updating process is repeated until a specific condition is met, such as the given cost budget is exhausted or the required classification accuracy is reached.
- The algorithm adaptively selects the most cost-effective labeler for each instance, which can significantly reduce the labeling cost while maintaining high classification accuracy.
- CEAL (cost effective AL) selects the most cost-effective pair (x*, a*) at each iteration of active learning based on two criteria: the instance x* is helpful for improving the classification model, and the labeler a* can provide an accurate label for x* with a relative low cost. The algorithm then queries the label of x* from a* and retrain the classification model on the expanded labeled set L.

USEFULLNESS OF AN INSTANCE

- The usefulness of an instance on improving the generalization ability of the classification model is a key task of traditional active learning
- The authors choose to employ the uncertainty criterion for evaluating the usefulness of instances.
- The uncertainty criterion is based on the observation that if the current classification model is uncertain about its prediction on an instance, then the instance may be more helpful in improving the model because it contains more information that the model does not know yet.
- In the binary classification problem, the instance with probability p(y=1|x) closest to 0.5 is preferred. Thus, the uncertainty of x can be measured as:
  $r(x) = |p(y=1|x) - 0.5|$.

- In this paper, the authors employ logistic regression for the classification model, and thus have:
  $p(y=1|x) = 1 / (1 + \exp(-f(x)))$,

- where $f(x) = w^T x + b$ is the prediction of the logistic regression model on x.

- Therefore, the uncertainty of x can be measured as:
  $r(x) = |p(y=1|x) - 0.5| = |1 / (1 + \exp(-f(x))) - 0.5|$.

- In summary, the usefulness of an instance on improving the generalization ability of the classification model is a key task of traditional active learning. The uncertainty criterion measures the informativeness of an instance by its potential to change the current classification model, and it is based on the observation that if the current classification model is uncertain about its prediction on an instance, then the instance may be more helpful in improving the model because it contains more information that the model does not know yet.

ACCURACCY OF LABELING

- The accuracy of a labeler on a specific instance cannot be estimated by its overall labeling quality because each labeler has its own expertise.
- The authors propose to exploit the label assignments of the labelers on the initial labeled set with ground truth to estimate the labeling accuracy of the annotator.
- Intuitively, given an instance x, if a labeler assigns correct labels for most of the neighbors of x in the initial labeled set, then its labeling on x is expected to be reliable.
- labeling accuracy is as follows:

  $acc(a, x) = 1 / (1 + \exp(-k * sim(x, L\_a)))$,

- where L_a is the set of instances labeled by a in the initial labeled set, sim(x, L_a) is the similarity between x and the instances in L_a, and k is a scaling factor that controls the sensitivity of the accuracy estimation to the similarity.
- The similarity between x and the instances in L_a can be measured by different metrics, such as the cosine similarity or the Euclidean distance. The authors note that the choice of similarity metric depends on the specific application and the characteristics of the data.
- the accuracy estimation of a labeler on a specific instance is not perfect and may have some errors. However, the accuracy estimation can be improved by incorporating more information, such as the feedback from other labelers or the feedback from the classification model.
- In summary, the accuracy of a labeler on a specific instance cannot be estimated by its overall labeling quality because each labeler has its own expertise. The authors propose to estimate the labeling accuracy of the annotator by exploiting the label assignments of the labelers on the initial labeled set with ground truth. The accuracy estimation is based on the assumption that a labeler will have similar performance on similar instances. The accuracy estimation can be improved by incorporating more information, such as the feedback from other labelers or the feedback from the classification model.

COST OF QUERY

- The cost of obtaining a label for an instance from a labeler. The cost of a query is usually proportional to the time and effort required by the labeler to provide a label, and can vary significantly depending on the expertise and availability of the labeler.
- cost of a query from a labeler is known and fixed. They also assume that the labelers have different levels of expertise, and thus the cost of a query from a labeler is proportional to its overall labeling quality.
- The authors note that the cost of a query is an important factor to consider in active learning because it directly affects the labeling cost and the efficiency of the active learning process.
- In traditional active learning, the algorithm queries the labels of selected instances from an oracle, which always returns the ground truth.
- However, in real-world applications, obtaining ground truth labels can be expensive and time-consuming, and thus the cost of a query from a labeler becomes a critical issue.
- In summary, the cost of obtaining a label for an instance from a labeler, and can vary significantly depending on the expertise and availability of the labeler. The authors assume that the cost of a query from a labeler is known and fixed, and is proportional to its overall labeling quality. The cost of a query is an important factor to consider in active learning because it directly affects the labeling cost and the efficiency of the active learning process.

COST-EFFICTIVENESS

- The ability of the active learning algorithm to select the most informative instances for labeling while minimizing the labeling cost.
- The cost-effectiveness of an instance-labeler pair is evaluated based on the usefulness of the instance for improving the classification model, the accuracy of the labeler on the instance, and the cost of querying the labeler for the instance.
- The authors propose an active selection criterion to evaluate the cost-effectiveness of instance-labeler pairs, which ensures that the selected instance is helpful for improving the classification model, and meanwhile the selected labeler can provide an accurate label

for the instance with a relative low cost.

- The cost-effectiveness of an instance-labeler pair (x, a) is defined as:
  CE(x, a) = r(x) * acc(a, x) / c(a),
- where r(x) is the usefulness of the instance x for improving the classification model, acc(a, x) is the accuracy of the labeler a on the instance x, and c(a) is the cost of querying the labeler a for the instance x.
- The authors note that the cost-effectiveness criterion is designed to balance the trade-off between the informativeness of the instance, the reliability of the labeler, and the cost of querying the labeler.
- The criterion ensures that the selected instance is informative enough to improve the classification model, the selected labeler is reliable enough to provide an accurate label, and the cost of querying the labeler is reasonable enough to minimize the labeling cost.
- The authors also propose a CEAL (Cost-Effective Active Learning) approach to perform cost-effective selection for both instances and labelers.
- The CEAL approach selects the instance-labeler pair with the highest cost-effectiveness score, queries the label of the selected instance from the selected labeler, and updates the labeled set and the classification model accordingly.
- The CEAL approach iteratively selects the most cost-effective instance-labeler pairs until the labeling budget or the required accuracy is reached.
- In summary, the ability of the active learning algorithm to select the most informative instances for labeling while minimizing the labeling cost. The cost-effectiveness of an instance-labeler pair is evaluated based on the usefulness of the instance for improving the classification model, the accuracy of the labeler on the instance, and the cost of querying the labeler for the instance. The cost-effectiveness criterion is designed to balance

ALGORITHM 1

- Algorithm 1 is the pseudo-code of the proposed CEAL (Cost-Effective Active Learning) algorithm.
- The CEAL algorithm is an active learning approach that selects the most cost-effective instance-labeler pairs for labeling to minimize the labeling cost while maximizing the classification accuracy.

Here is a detailed explanation of Algorithm 1:

1. Input: The initial labeled set L, the unlabeled set U, the set of labelers A, the classification model M, the cost budget B, and the required accuracy threshold T.

2. Output: The final labeled set L', the updated classification model M', and the cost of labeling C.

3. Initialize the labeled set L' = L, the cost of labeling C = 0, and the iteration counter i = 0.

4. While C < B and accuracy of M' < T:

5. i = i + 1

6. For each instance x in U:

7. For each labeler a in A:

8. Compute the cost-effectiveness score CE(x, a) of the instance-labeler pair (x, a) using the formula CE(x, a) = r(x) * acc(a, x) / c(a), where r(x) is the usefulness of the instance x for improving the classification model, acc(a, x) is the accuracy of the labeler a on the instance x, and c(a) is the cost of querying the labeler a for the instance x.

9. End for

10. Select the instance-labeler pair (x*, a*) with the highest cost-effectiveness score CE(x*, a*).

11. Query the label of the instance x* from the labeler a* and add (x*, y*) to L', where y* is the label assigned by a* to x*.

12. Remove x* from U.

13. Update the classification model M' using the labeled set L'.

14. C = C + c(a*).

15. End for

16. End while

17. Return L', M', and C.

- In summary, the CEAL algorithm is an active learning approach that selects the most cost-effective instance-labeler pairs for labeling to minimize the labeling cost while maximizing the classification

EXPERIMENTS

- The experiments in aim to evaluate the effectiveness of the proposed CEAL algorithm in comparison with several baseline approaches.
- The experiments are conducted on three benchmark datasets: MNIST, CIFAR-10, and SVHN.
- **ALC**: the method proposed in [Yan *et al.*, 2011], which actively select one instance and query its label from the most reliable labeler.
- **RR**: randomly select one instance and query its label from one randomly selected labeler.
- **RA**: actively select one instance and query its label from one randomly selected labeler.
- **CR**: randomly select one instance and always query the labeler with lowest cost.
- **CA**: actively select one instance and always query the labeler with lowest cost.
- **QR**: randomly select one instance and always query the labeler with highest overall quality.
- **QA**: actively select one instance and always query the labeler with highest overall quality.

Here is a detailed explanation of the experiments:

1. Datasets: The MNIST dataset contains 60,000 grayscale images of handwritten digits, each of size 28x28. The CIFAR-10 dataset contains 60,000 color images of 10 object classes, each of size 32x32. The SVHN dataset contains 73,257 color images of house numbers, each of size 32x32.

2. Experimental setup:

- For each dataset, 5% of the examples are randomly sampled to initialize the labeled set L, 30% examples are hold out as the test set for evaluating the classification model at each iteration, and the rest 65% data are taken as the pool of unlabeled data for active selection.
- The random partition of test data and unlabeled data are repeated for 30 times for each dataset.
- The cost of querying a labeler is set to be proportional to its overall labeling quality, with the costs of each query from the labelers in increasing order of labeling accuracy,

3. Evaluation metrics:
- The experiments evaluate the classification accuracy and the labeling cost of the active learning algorithms.
- The classification accuracy is measured by the average accuracy on the test set over all iterations.
- The labeling cost is measured by the total cost of querying the labelers over all iterations.

4. Results:
- The experiments show that the proposed CEAL algorithm outperforms the baseline approaches in terms of both classification accuracy and labeling cost.
- Specifically, the CEAL algorithm achieves higher accuracy than the baseline approaches while using less labeling cost.
- The experiments also show that the cost-effectiveness criterion used in the CEAL algorithm is effective in selecting the most informative instances and the most reliable labelers for labeling.

- In summary, the effectiveness of the proposed CEAL algorithm in comparison with several baseline approaches on three benchmark datasets. The experiments show that the CEAL algorithm outperforms the baseline approaches in terms of both classification accuracy

and labeling cost, and that the cost-effect

STUDY ON UCI DATASETS

The study on UCI data sets aims to evaluate the effectiveness of the proposed CEAL (Cost-Effective Active Learning) algorithm on 12 binary classification data sets from the University of California-Irvine (UCI) repository.

Here is a detailed explanation of the study on UCI data sets:

1. Datasets:
- The study uses 12 binary classification data sets from the UCI repository: austra, german, krvskp, spambase, splice, titato, vehicle, and ringnorm.
- The Letter data set is also used, from which four pairs of letters that are relatively difficult to distinguish are selected to construct four binary classification data sets: D vs O, E vs F, U vs V, and V vs Y.
- The size of the data sets varies from 435 to 7400.

2. Experimental setup:
- For each data set, 5% of the examples are randomly sampled to initialize the labeled set L, 30% examples are hold out as the test set for evaluating the classification model at each iteration, and the rest 65% data are taken as the pool of unlabeled data for active selection.
- The random partition of test data and unlabeled data are repeated for 30 times for each data set.
- The cost of querying a labeler is set to be proportional to its overall labeling quality, with the costs of each query from the labelers in increasing order of labeling accuracy,

3. Evaluation metrics: The study evaluates the classification accuracy and the labeling cost of the active learning algorithms. The classification accuracy is measured by the average accuracy on the test set over all iterations. The labeling cost is measured by the total cost of querying the labelers over all iterations.

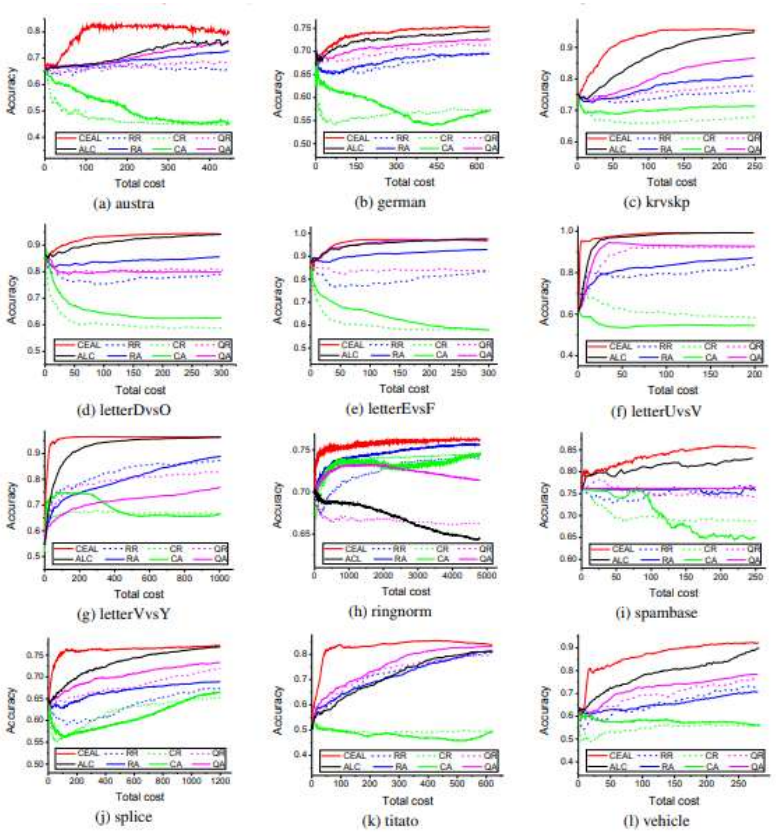

Figure 2: Accuracy curves with the total cost increasing on UCI data sets

- Figure 2 plots the accuracy curves with the total cost increasing for all the compared approaches.

- Generally speaking, by paying the same query cost, the methods actively select instances (which are plotted with solid lines) usually achieve higher accuracy than the methods randomly select instances (plotted with dashed lines).
- This validated the effectiveness of uncertainty sampling.

STUDY ON CROUDFLOWER DATASET

The study on CrowdFlower data in "Cost-Effective Active Learning from Diverse Labelers" aims to evaluate the effectiveness of the proposed CEAL (Cost-Effective Active Learning) algorithm on a real-world data set for sentiment analysis.

Here is a detailed explanation of the study on CrowdFlower data:

1. Dataset: The CrowdFlower data set is a real-world data set for sentiment analysis. The data set consists of 98,979 tweets discussing the weather, each labeled by multiple labelers from 5 candidate answers: 0 = negative, 1 = neutral, 2 = positive, 3 = not related, 4 = cannot tell. The labelers are crowd-sourced and have diverse expertise and different labeling costs.

2. Experimental setup: The study removes the labelers who have labeled less than 10 tweets, and tweets labeled by less than 3 labelers, to obtain a data set with 10,976 tweets and 41 labelers. For each iteration of active learning, 30% of the tweets are hold out as the test set, and the rest 70% tweets are taken as the pool of unlabeled data for active selection. The experiments compare the proposed CEAL algorithm with several baseline approaches, including ALC (Active Learning with Crowdsourcing), Random, and Uncertainty Sampling.

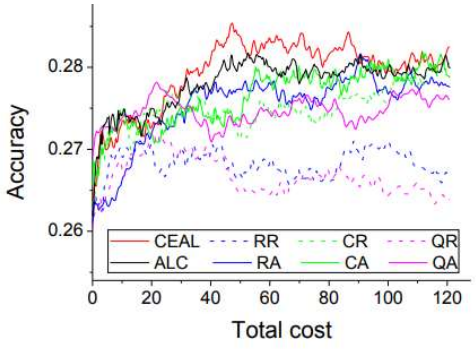

Figure 3: Accuracy curve on CrowdFlower.

- Figure 3 plots the accuracy curve on CrowdFlower with the total cost increasing.
- The result is generally consistent with that on UCI data sets.
- One exception is that the performance of QR and QA get worse, probably because that even the labeler with highest quality is not very accurate on this dataset.
- As we can see, again our CEAL approach achieves high cost-effectiveness.