Cost-Effective Active Learning from Diverse Labelers*

Sheng-Jun Huang[1,3], **Jia-Lve Chen**[2,3], **Xin Mu**[2,3] and **Zhi-Hua Zhou**[2,3]
[1]College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics
[2]National Key Laboratory for Novel Software Technology, Nanjing University
[3]Collaborative Innovation Center of Novel Software Technology and Industrialization
huangsj@nuaa.edu.cn {chenjl, mux, zhouzh}@lamda.nju.edu.cn

Abstract

In traditional active learning, there is only one labeler that always returns the ground truth of queried labels. However, in many applications, multiple labelers are available to offer diverse qualities of labeling with different costs. In this paper, we perform active selection on both instances and labelers, aiming to improve the classification model most with the lowest cost. While the cost of a labeler is proportional to its overall labeling quality, we also observe that different labelers usually have diverse expertise, and thus it is likely that labelers with a low overall quality can provide accurate labels on some specific instances. Based on this fact, we propose a novel 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. Experiments on both UCI and real crowdsourcing data sets demonstrate the superiority of our proposed approach on selecting cost-effective queries.

1 Introduction

In real-world applications, label acquisition for the abundant of unlabeled data is usually expensive because it requires the participation of human experts. As a result, training an accurate model with as few labeled data as possible has attracted a lot of research interests. One leading approach to this goal is active learning, which reduces the labeling cost by querying only the most valuable instances for their class assignments [Settles, 2009].

In traditional active learning, the algorithm iteratively queries the labels of selected instances from an oracle, which always returns the ground truth. Under this setting with a unique labeler, the labeling cost is measured by the number of queries, and the goal is to train an accurate model with fewest queries. Active learning under this simplified setting has been extensively studied, however, it is not in conformity with the reality. In many real tasks, we usually have multiple labelers offering different qualities of labeling, and need to decide which of them should be queried for the label of the selected instance. The overall quality of labelers can be evaluated based on their previous performance. A labeler who made fewer mistakes receives a higher quality score, and accordingly requires a higher cost for each query. For example, in the task of breast cancer diagnosis from medical images, doctors cannot guarantee 100% accuracy for their classifications, and in general, experienced experts are more likely to make accurate diagnosis than junior doctors. Also, in the crowdsourcing annotation services, such as Amazon's Mechanical Turk, thousands of labelers are available on the Internet to provide noisy annotations, and those labelers who are more reliable tend to receive higher payment.

In recent years, learning with multiple labelers has attracted more and more research interests. Most studies focus on obtaining an accurate label from multiple noisy labelers [Donmez *et al.*, 2010; Sheng *et al.*, 2008; Whitehill *et al.*, 2009; Yan *et al.*, 2014; Raykar *et al.*, 2010; Ipeirotis *et al.*, 2014; Lin *et al.*, 2014], ignoring the active selection of instances and labelers. Several approaches are proposed for active learning with multiple labelers [Zheng *et al.*, 2010; Zhao *et al.*, 2011; Zhang *et al.*, 2015]. However, they usually neglect the diverse expertise of different labelers, and globally select the same labeler for all instances. Recently, there are a few attempts to exploit the expertise of labelers for better annotation [Yan *et al.*, 2011; Fang *et al.*, 2014; Dekel *et al.*, 2012; Yan *et al.*, 2012; Ambati *et al.*, 2010; Donmez *et al.*, 2009]. However, they simply assume an identical cost for all labelers, and may get an expensive solution.

One significant difference between our work and existing active learning approaches is that we emphasize the diversity of labelers on both their expertise and query cost. In real tasks, labelers can be different in age, gender, interests and so on, and thus they may have very diverse expertise. As a result, different labelers are good at labeling different instances, which further indicates that a labeler with very low overall quality can still make accurate labeling on some specific instances. For example, in image annotation tasks, a child labeler may make many mistakes on identifying the location of landscape photos, but can easily assign accurate labels to

*This research was supported by the 973 Program (2014CB340501), JiangsuSF (BK20150754) and NSFC (61333014, 61503182). This research was started when S.-J. Huang was in the LAMDA Group, Nanjing University.

cartoon images; or in evaluation of products, male customers must be more confident on rating a shaver than a skirt. While the overall quality of labeling directly decides the cost of a labeler, it is not necessarily related to the accuracy of the label on a specific instance. And based on this fact, for a given instance, we can hopefully find a right labeler which provides accurate yet cheap annotations.

In this paper, we propose a novel criterion for active selection on both instances and labelers. The criterion evaluates the cost-effectiveness of instance-labeler pairs, and ensures that on one hand, the selected instance is helpful on improving the classification model, and on the other hand, the selected labeler can assign an accurate label for the instance with a low cost. Extensive experimental results validated the superiority of our proposed approach.

The rest of this paper is organized as follows: we first review some related work in Section 2, and then introduce the proposed approach in Section 3. Section 4 presents the experiments, followed by the conclusion in Section 5.

2 Related Work

Active learning has been well studied during the past years. Many algorithms are proposed to design different criteria for query selection such that the queried examples are most helpful for improving the classification model. Based on the query criteria used, active learning methods can be roughly categorized into three groups: the methods querying most informative instances (e.g., uncertainty sampling [Balcan *et al.*, 2007; Guo and Schuurmans, 2008] and expected error reduction based sampling [Roy and Mccallum, 2001]); the methods querying most representative instances (e.g., clustering based sampling [Dasgupta and Hsu, 2008; Chattopadhyay *et al.*, 2012] and density based sampling [Zhu *et al.*, 2010]); and the methods simultaneously consider informativeness and representativeness [Huang *et al.*, 2014; Wang and Ye, 2013].

The standard active learning assumes that there is only one oracle which never makes mistake on its labeling. In real-world tasks, however, multiple noisy labelers are available. Many methods have been proposed to learn from the multiple noisy labelers. Some of them try to model the reliability of labelers, and then select the best ones for labeling all the instances [Donmez *et al.*, 2010]. Some other approaches try to find a best combination of noisy labels given by all labelers [Sheng *et al.*, 2008; Whitehill *et al.*, 2009]. These methods, however, do not consider the active selection of instances. The study in [Giacinto and Roli, 2000] is partially related to dynamic classifier selection in ensemble methods, which dynamically selects one classifier for each test instance.

There are some studies on active learning with multiple noisy labelers. Zhao et. al. [2011] combine uncertainty and inconsistency measures to actively select the most important instances for relabeling. Although this approach assumes the existence of multiple noisy labelers, it only designs the criterion for instance selection, and does not select labelers. The approach in [Zheng *et al.*, 2010] obtains the label from a subset of labelers via majority voting. The labelers are globally selected for all instances, which may result in unnecessary high cost because it neglects the expertise of different labelers. There are a few studies trying to select labelers with matching expertise on the specific instance to be labeled [Dekel *et al.*, 2012; Yan *et al.*, 2012; Donmez *et al.*, 2009; Ipeirotis *et al.*, 2014]. For example, Yan et. al. [2011] propose a probabilistic model to measure the accuracy of each labeler, and select the most confident labeler for each queried instance. Fang et. al. [2014] propose to transfer knowledge from auxiliary domains for estimating labelers expertise. Ambati et. al. [2010] apply the active selection of both instances and labelers to the task of machine translation. A common shortcoming of these algorithms is that they do not consider the difference on the costs of multiple labelers, and thus may get an accurate yet expensive solution. The methods proposed in [Donmez and Carbonell, 2008] and [Zhang and Chaudhuri, 2015] consider a simple case where two labelers exist: one perfect labeler which always returns the ground truth, and one fallible labeler which make mistakes with a probability. Although they assume different costs for the two labelers, the oversimple setting limits its application.

3 The Proposed Algorithm

We study the task of active learning from multiple labelers with diverse expertise and different costs. At each iteration of active learning, we select a cost-effective instance-labeler pair with following properties: 1) the selected instance is useful on improving the classification model; 2) the selected labeler can assign an accurate class label for the instance; and 3) the cost of the selected labeler is low.

We start with a simple example to illustrate our basic idea. Figure 1 shows an extreme case of three labelers with different expertise. The data examples are partitioned into three groups: A, B and C, with decreasing sizes. Labeler 1 is good at labeling examples in group A, but fails on groups B and C. Labeler 2 works well on groups A and B, while labeler 3 favors groups B and C. Obviously, labeler 2 achieves the best accuracy of labeling on the whole data set, and deserves the highest query cost. In contrast, labeler 3's overall labeling quality is the worst, and accordingly receives the lowest cost. While the overall quality of labeling decides the cost of each query from a labeler, it does not necessarily implies the accuracy on a specific instance. For the instances in group B, both labeler 2 and labeler 3 can assign correct labels, and then labeler 3 with lower cost is preferred, though its overall quality is lower. This observation suggests that globally selecting one labeler for all instances may result in unnecessary high cost, and one better choice is to adaptively select the most cost-effective labeler for each instance.

Assume that we have a small labeled set $L = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_l}$ with n_l examples, where y_i is the ground truth label of the i-th instance $\mathbf{x}_i$, and a large pool of unlabeled data $U = \{\mathbf{x}_j\}_{j=n_l+1}^{n_l+n_u}$ with n_u instances, typically $n_l \ll n_u$. There is a set of candidate labelers $A = \{a_1, \cdots, a_m\}$ with all m labelers offering different qualities of labeling with different prices. We denote by $\hat{y}_{ij}$ the label given by a_i for $\mathbf{x}_j$.

At each iteration of active learning, the algorithm selects an instance-labeler pair $(\mathbf{x}^*, a^*)$, and queries the label of $\mathbf{x}^*$ from a^*, where the selection of both the instance and labeler

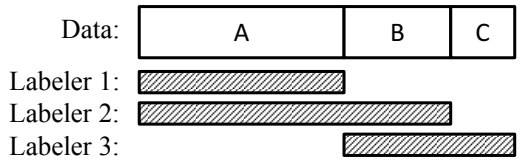

Figure 1: Multiple labelers with diverse expertise. The textured boxes indicate on which parts of data each labeler can give accurate labels.

is based on a evaluation function Q:

$$(\mathbf{x}^*, a^*) = \arg\max_{\mathbf{x}\in U,\ a\in A} Q(\mathbf{x}, a).$$

Then the task is to design the evaluation function $Q(\mathbf{x}, a)$ to measure the cost-effectiveness of the pair $(\mathbf{x}, a)$. To make it clear how the designed criterion satisfies the previously listed properties, we first separately introduce the usefulness of an instance, the accuracy of the labeling and the cost of the query, and then propose the evaluation function by combining them together.

3.1 Usefulness of the Instance

Evaluating the usefulness of an instance on improving the generalization ability of the classification model is the key task of traditional active learning, and has been well studied. Among the existing criteria for evaluating the usefulness of instances, we choose to employ the uncertainty, which is most widely used. If the current classification model is uncertain about its prediction on an instance, then the instance may be more helpful on improving the model because it contains more information that the model does not know yet.

In the binary classification problem, the instance with probability $p(y = 1|\mathbf{x})$ closest to 0.5 is preferred. Thus the uncertainty of $\mathbf{x}$ can be measured as:

$$r(\mathbf{x}) = |p(y = 1|\mathbf{x}) - 0.5|. \quad (1)$$

In this paper, we employ logistic regression for the classification model, and thus have

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-f(\mathbf{x}))},$$

where $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ is the prediction of the logistic regression model on $\mathbf{x}$.

To be more general for multi-class setting, the uncertainty $r(\mathbf{x})$ can be rewritten as:

$$r(\mathbf{x}) = 1 - \max_{y\in\mathcal{Y}} p(y|\mathbf{x}), \quad (2)$$

where $p(y|\mathbf{x})$ is the probability of that $\mathbf{x}$ belongs to class y, and $\mathcal{Y}$ is the set of all possible classes.

3.2 Accuracy of the Labeling

As discussed before, the accuracy of a labeler on a specific instance cannot be estimated by its overall labeling quality. Because each labeler has its own expertise, it is reasonable to assume that a labeler will have similar performance on similar instances. We thus try to exploit the label assignments of the labelers on the initial labeled set with ground truth. Intuitively, given an instance $\mathbf{x}$, if a labeler assigns correct labels for most of the neighbors of $\mathbf{x}$ in the initial labeled set, then its labeling on $\mathbf{x}$ is expected to be reliable. Formally ,we estimate the labeling accuracy of the annotator a_i on instance $\mathbf{x}_j$ by:

$$q_i(\mathbf{x}_j) = \frac{1}{t} \sum_{\mathbf{x}_k\in N(\mathbf{x}_j,t)} S(\mathbf{x}_j, \mathbf{x}_k) I[y_k == \hat{y}_{ik}], \quad (3)$$

where $N(\mathbf{x}_j, t)$ returns a set consists of t nearest neighbors of $\mathbf{x}_j$ in the initial labeled set, $S(\mathbf{x}_j, \mathbf{x}_k)$ measures the similarity between $\mathbf{x}_j$ and $\mathbf{x}_k$, and $I[\cdot]$ is the indicator function which returns 1 if the argument is true and 0 otherwise. In our experiments, we simply select the neighbors based on Euclidean distance. Obviously, among the t neighbors, the instances which are more similar to $\mathbf{x}_j$ contribute more to the estimation of $q_i(\mathbf{x}_j)$.

3.3 Cost of the Query

Typically, if a labeler provides high quality of labeling, it requires a high price for answering each query. So the cost of each query from a labeler may be decided by the accuracy of its previous labeling. We denote by c_i the cost of querying one label from labeler a_i, and define it as

$$c_i = g(\frac{1}{n_l} \sum_{j=1}^{n_l} I[y_j == \hat{y}_{ij}]), \quad (4)$$

where $\frac{1}{n_l} \sum_{j=1}^{n_l} I[y_j == \hat{y}_{ij}]$ calculates the overall accuracy of labeler a_i on L, and $g(\cdot)$ is the function describing the relation between the cost and accuracy. For simplicity, one can implement it as any increasing function, e.g. $g(z) = z$.

3.4 Cost-Effectiveness

Based on the previous definitions, we then try to propose the evaluation function Q for active selection. To select a cost-effective instance-labeler pair $(\mathbf{x}_j, a_i)$, we want the *usefulness* $r(\mathbf{x}_j)$ to be large, the *accuracy* $q_i(\mathbf{x}_j)$ to be high, and at the same time the *cost* c_i to be low. A straightforward criterion for estimating the cost-effectiveness of the pair $(\mathbf{x}_j, a_i)$ can be defined as Eq. 5.

$$Q(\mathbf{x}_j, a_i) = \frac{q_i(\mathbf{x}_j) \cdot r(\mathbf{x}_j)}{c_i}. \quad (5)$$

And the most cost-effective instance-labeler pair is selected as:

$$(\mathbf{x}^*, a^*) = \arg\max_{\mathbf{x}_j\in U, a_i\in A} Q(\mathbf{x}_j, a_i). \quad (6)$$

At last, let us revisit the three properties of the selection we discussed at the beginning of this section. Obviously, an instance-labeler pair violating either of the three properties will receive a small score of Eq. 5, and thus the proposed active selection is cost-effective.

It is worth noting that though the simple implementation of the proposed approach achieves good results to demonstrate

Algorithm 1 The CEAL Algorithm

Input:
L : a small set of labeled data
U : the pool of unlabeled data for active selection
$\hat{Y}$: the labels given by all labelers on L
Initialize:
Calculate the the cost for all the labelers as Eq. 4
repeat
for each instance $\mathbf{x}_j \in U$ and each labeler a_i
Calculate the uncertainty for $\mathbf{x}_j$ according to Eq. 2
Predict the accuracy of a_i on $\mathbf{x}_j$ according to Eq. 3
Calculate the cost-effectiveness for $(\mathbf{x}_j, a_i)$ as Eq. 5
end for
Select the pair $(\mathbf{x}^*, a^*)$ as Eq. 6
Query the label of $\mathbf{x}^*$ from a^*, denoted by $\hat{y}^*$
$L = L \cup (\mathbf{x}^*, \hat{y}^*)$; $U = U \setminus \mathbf{x}^*$
Train the classification model on L, and evaluate the model on the test set
until the cost budget or the required accuracy is reached

our idea, we can have other choices for each of the three components in Eq. 5. For example, the *usefulness* $r(\mathbf{x}_j)$ can be designed as in [Huang *et al.*, 2014] to consider both informativeness and representativeness of the instances; the labeling *accuracy* $q_i(\mathbf{x}_j)$ can be estimated by a individual model for each labeler; and the *cost* c_i can be specified by users according to actual situations.

The pseudo-code of the proposed algorithm, termed CEAL (Cost-Effective Active Learning), is summarized in Algorithm 1. At each iteration of active learning, the algorithm selects the most cost-effective pair $(\mathbf{x}^*, a^*)$, and queries the label of $\mathbf{x}^*$ from a^*. Then $\mathbf{x}^*$ is removed from the unlabeled set U, and is added into L along with its queried label $\hat{y}^*$ from a^*. After that, the classification model is retrained on the expanded labeled set L. At last, the updated model is evaluated on a hold out test set. This active querying and model updating process is repeated until meets a specific condition, e.g., the given cost budget is exhausted or the required classification accuracy is reached.

4 Experiments

We compare CEAL with the following baseline approaches:

- **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.

Table 1: The accuracy of each labeler on L for UCI data sets

data	accuracy of labelers on L				
	a_1	a_2	a_3	a_4	a_5
austra	0.943	0.771	0.743	0.686	0.571
german	0.920	0.860	0.800	0.740	0.640
krvskp	0.881	0.875	0.719	0.694	0.631
letterDvsO	0.808	0.744	0.679	0.641	0.615
letterEvsF	0.896	0.766	0.740	0.714	0.571
letterUvsV	0.937	0.747	0.759	0.734	0.557
letterVvsY	0.923	0.744	0.756	0.731	0.538
spambase	0.917	0.878	0.804	0.696	0.609
splice	0.899	0.711	0.704	0.686	0.509
titato	0.875	0.750	0.729	0.708	0.521
vehicle	0.818	0.773	0.727	0.636	0.545
ringnorm	0.824	0.819	0.781	0.770	0.624

- **QA**: actively select one instance and always query the labeler with highest overall quality.

For all the compared approaches, uncertainty sampling is used for instance selection, where the uncertainty is measured based on the logistic regression model trained on the labeled data. We also evaluate the performance on test data by the logistic regression model implemented with LIBLINEAR [Fan *et al.*, 2008] with default parameters.

4.1 Study on UCI data sets

We first perform the experimental study on 12 data sets from the University of California-Irvine (UCI) repository [Bache and Lichman, 2013]: *austra*, *german*, *krvskp*, *spambase*, *splice*, *titato*, *vehicle* and *ringnorm*. *Letter* is a multi-class data set, from which we select four pairs of letters that are relatively difficult to distinguish, i.e., *D vs O*, *E vs F*, *U vs V*, *V vs Y*, and construct a binary class data set for each pair. The size of the data sets varies from 435 to 7400.

We assume that there are in all 5 labelers offering different qualities of labeling: a_1, a_2, a_3, a_4 and a_5. Table 1 presents the accuracy of labeling on the initial labeled set L given by each labeler. In increasing order of labeling accuracy, we set the costs of each query from the labelers as 1 to 5, respectively. For each data set, 5% of the examples are 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. And we report the average results over the 30 runs of experiments.

Figure 2 plots the accuracy curves with the total cost increasing for all the compared approaches. Note that we did not plot the full curve because the performances of most methods have converged after a number of queries. 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. As expected, querying all instances from labeler with minimal cost leads to the worst

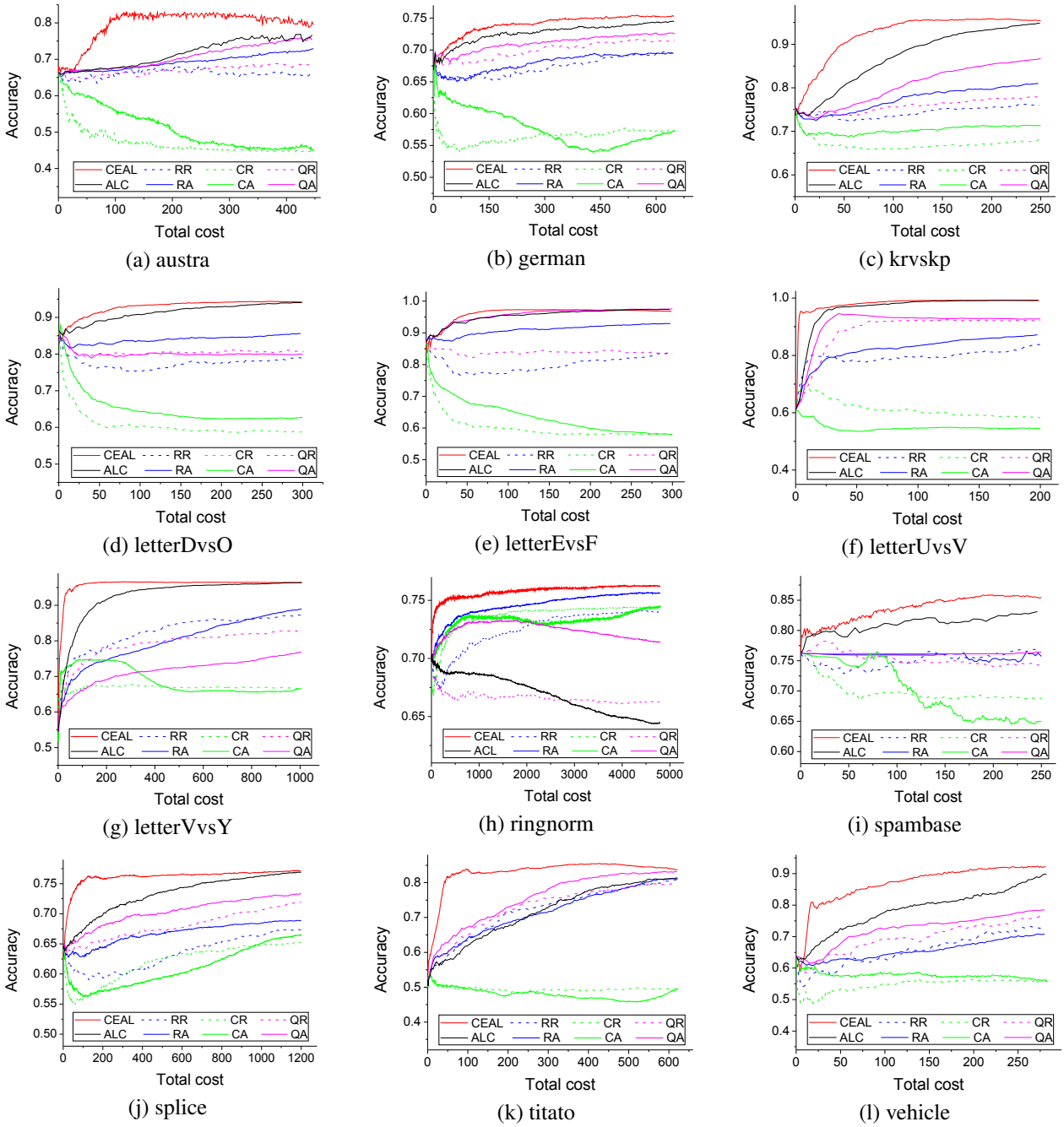

Figure 2: Accuracy curves with the total cost increasing on UCI data sets

performance since it receives too many noisy labels. Querying the labeler with highest overall quality may get most instances correctly labeled, but also results in a high cost. Random selection of labelers receives a middle performance. We observe that ALC achieves decent performance on most of the data sets, and fails on the others. Because ALC selects the most confident labeler for each instance, but does not consider cost, it is likely to get accurate labels with high cost, making it less cost-effective. At last, our proposed approach CEAL consistently achieves higher accuracy than other methods with the same query cost.

After the comparison on the cost-effectiveness, we further examine the difference between the compared methods on labeler selection. Table 2 presents the average cost of each query for all the compared methods. Because the counterparts that select instances actively or randomly have similar results, we do not report the results for RR, CR and QR. First, the cost of CA and QA exactly equals 1 and 5, respectively,

Table 2: The average cost of each query for all the compared approaches

	Ours	RA	CA	QA	ALC
austra	1.554	3.004	1.000	5.000	4.037
german	1.477	3.019	1.000	5.000	3.489
krvskp	1.447	3.000	1.000	5.000	3.448
letterDvsO	1.496	2.986	1.000	5.000	4.030
letterEvsF	1.575	3.012	1.000	5.000	4.429
letterUvsV	1.470	2.987	1.000	5.000	3.522
letterVvsY	1.417	2.992	1.000	5.000	4.114
spambase	1.390	2.997	1.000	5.000	4.158
splice	1.459	3.000	1.000	5.000	3.191
titato	1.463	3.009	1.000	5.000	3.090
vehicle	1.657	2.989	1.000	5.000	2.219
ringnorm	1.295	3.000	1.000	3.000	3.304
average	1.475	3.000	1.000	4.833	3.586

Table 3: The accuracy of queried labels for all approaches

	Ours	RA	CA	QA	ALC
austra	0.928	0.735	0.557	0.934	0.992
german	0.933	0.791	0.654	0.920	0.999
krvskp	0.933	0.757	0.641	0.895	0.994
letterDvsO	0.909	0.691	0.604	0.792	0.992
letterEvsF	0.950	0.745	0.575	0.898	0.968
letterUvsV	0.976	0.742	0.536	0.933	0.996
letterVvsY	0.963	0.744	0.557	0.907	0.990
spambase	0.940	0.784	0.617	0.905	0.995
splice	0.878	0.715	0.526	0.894	0.883
titato	0.823	0.716	0.523	0.876	0.821
vehicle	0.890	0.687	0.532	0.824	0.860
ringnorm	0.736	0.759	0.620	0.816	0.772
average	0.905	0.739	0.578	0.883	0.939

because they always query the labels from the same labeler. The cost of our proposed method CEAL is much lower than ALC and RA by actively selecting the cost-effective labeler.

Similarly, we present in Table 3 the accuracy of queried labels for all approaches, i.e., the percentage of queried instances that have been correctly labeled. CA selects the labeler with lowest cost, and thus has an accuracy close to the accuracy of a_5 in Table 1. Similarly, QA has the accuracy of the most accurate labeler. It is not surprising to observe that ALC achieves a high accuracy because it selects to query the most confident labeler for each instance. The advantage of ALC to QA validates that adaptively selecting a labeler for each instance is superior to the global selection for all instances. Our proposed approach CEAL also achieves a high accuracy, only slightly worse than ALC. Noticing the significant lower cost of our approach, it is obvious that our approach CEAL selects the most cost-effective queries.

4.2 Study on CrowdFlower data

CrowdFlower is a data set for sentiment analysis. The crowdsourced labelers are asked to judge the sentiment of tweets discussing the weather. The data set consists of 98979 tweets, each was labeled by multiple labelers from 5 candidate answers: $0 = negative$, $1 = neutral$, $2 = positive$, $3 = not\ related$, $4 = cannot\ tell$.

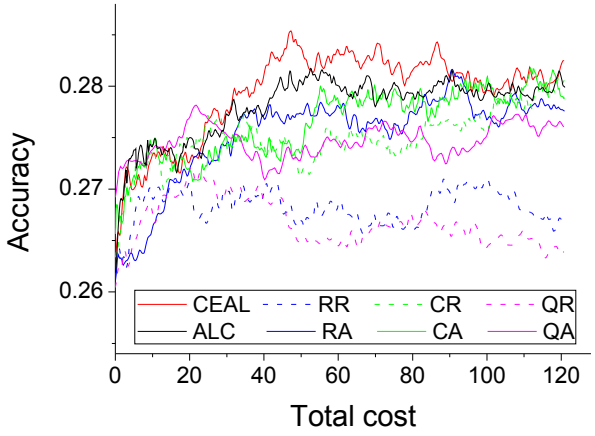

Figure 3: Accuracy curve on CrowdFlower.

By removing the labelers who have labeled less than 10 tweets, and tweets labeled by less than 3 labelers, we obtain a data set with 10976 tweets and 41 labelers. For each tweet, we extract a bag-of-words feature vector with 43395 dimensions. Among the tweets, only 269 tweets are given with the ground truth. We take 10 of them as the initial labeled set, and take the rest 259 examples as the test set for evaluating the classification model. The rest 10707 tweets are randomly partitioned into 10 folds. At each run of the experiment, we take one fold as the unlabeled pool of active learning. The average results over the 10 runs are recorded. In usual case, we estimate the cost of each labeler on the initial labeled set. However, there is too few examples with ground-truth. We instead estimate the cost on the test set, which is less than 5% of the whole dataset. Note that since we have a relative large number of labelers, the cost is set to equal the overall accuracy for each labeler, i.e., we set $g(z) = z$ in Eq. 4. The cost of the 41 labelers varies from 0.067 to 0.500.

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 data set. As we can see, again our CEAL approach achieves high cost-effectiveness.

5 Conclusion

This paper studies active learning under a novel setting, where multiple noisy labelers with diverse expertise and different labeling costs are available. We observe that a labeler with low quality of overall labeling can still assign accurate labels for some specific instances. Based on this fact, we propose the CEAL approach to perform cost-effective selection for both instances and labelers. Experimental results on both UCI and real data sets demonstrate that the proposed approach is able to achieve higher accuracy with lower query cost. While our current study assumes the cost of querying different instances from the same labeler is identical, we plan to incorporate instance-dependent query cost into our CEAL approach in the future.

References

[Ambati *et al.*, 2010] V. Ambati, S. Vogel, and J. G. Carbonell. Active learning and crowd-sourcing for machine translation. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2010.

[Bache and Lichman, 2013] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[Balcan *et al.*, 2007] M. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory*, pages 35–50, 2007.

[Chattopadhyay *et al.*, 2012] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye. Batch mode active sampling based on marginal probability distribution matching. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 741–749, 2012.

[Dasgupta and Hsu, 2008] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 208–215, 2008.

[Dekel *et al.*, 2012] O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *Journal of Machine Learning Research*, 13:2655–2697, 2012.

[Donmez and Carbonell, 2008] P. Donmez and J. G. Carbonell. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 619–628, 2008.

[Donmez *et al.*, 2009] P. Donmez, J. G. Carbonell, and J. G. Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2009.

[Donmez *et al.*, 2010] P. Donmez, J. G. Carbonell, and J. G. Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In *Proceedings of the SIAM International Conference on Data Mining*, pages 826–837, 2010.

[Fan *et al.*, 2008] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871–1874, 2008.

[Fang *et al.*, 2014] M. Fang, J. Yin, and D. Tao. Active learning for crowdsourcing using knowledge transfer. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1809–1815, 2014.

[Giacinto and Roli, 2000] G. Giacinto and F. Roli. Dynamic classifier selection. In *Proceedings of the 1st Workshop on Multiple Classifier Systems*, pages 177–189, 2000.

[Guo and Schuurmans, 2008] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In *Advances in Neural Information Processing Systems 21*, pages 593–600. 2008.

[Huang *et al.*, 2014] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (10):1936–1949, 2014.

[Ipeirotis *et al.*, 2014] Panagiotis G Ipeirotis, Foster Provost, Victor S Sheng, and Jing Wang. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441, 2014.

[Lin *et al.*, 2014] Christopher H Lin, Daniel S Weld, et al. To re (label), or not to re (label). In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[Raykar *et al.*, 2010] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[Roy and Mccallum, 2001] N. Roy and A. Mccallum. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the 18th International Conference on Machine Learning*, pages 441–448, 2001.

[Settles, 2009] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2009.

[Sheng *et al.*, 2008] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 614–622, 2008.

[Wang and Ye, 2013] Z. Wang and J. Ye. Querying discriminative and representative samples for batch mode active learning. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 158–166, 2013.

[Whitehill *et al.*, 2009] J. Whitehill, T. Wu, J. Bergsma, J. R Movellan, and P. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043. 2009.

[Yan *et al.*, 2011] Y. Yan, G. M Fung, R. Rosales, and J. Dy. Active learning from crowds. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1161–1168, 2011.

[Yan *et al.*, 2012] Y. Yan, R. Rosales, G. Fung, F. Farooq, B. Rao, and J. Dy. Active learning from multiple knowledge sources. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 1350–1357, 2012.

[Yan *et al.*, 2014] Y. Yan, R. Rosales, G. Fung, S. Ramanathan, and J. Dy. Learning from multiple annotators with varying expertise. *Machine Learning*, 95(3):291–327, 2014.

[Zhang and Chaudhuri, 2015] Chicheng Zhang and Kamalika Chaudhuri. Active learning from weak and strong labelers. In *Advances in Neural Information Processing Systems*, pages 703–711. 2015.

[Zhang *et al.*, 2015] J. Zhang, Z. Wu, and V. Shengs. Active learning with imbalanced multiple noisy labeling. *IEEE Transactions on Cybernetics*, 45(5):1081–1093, 2015.

[Zhao *et al.*, 2011] L. Zhao, G. Sukthankar, and R. Sukthankar. Incremental relabeling for active learning with noisy crowdsourced annotations. In *Proceedings of IEEE 3rd International Conference on Social Computing*, pages 728–733, 2011.

[Zheng *et al.*, 2010] Y. Zheng, S. Scott, and K. Deng. Active learning from multiple noisy labelers with varied costs. In *IEEE 10th International Conference on Data Mining*, pages 639–648, 2010.

[Zhu *et al.*, 2010] J. Zhu, H. Wang, B. Tsou, and M. Ma. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1323–1331, 2010.