Learning from Imbalanced Crowdsourced Labeled Data

Wentao Wang* Joseph Thekinen† Xiaorui Liu* Zitao Liu‡ Jiliang Tang*

Abstract

Crowdsourcing has proven to be a cost-effective way to meet the demands for massive labeled training data in supervised deep learning models. However, the obtained crowdsourced labels are often inconsistent and noisy due to cognitive and expertise differences among crowd workers. Existing approaches either infer latent true labels from noisy crowdsourced labels or learn a discriminative model directly from the crowdsourced labeled data, assuming the latent true label distribution is class-balanced. Unfortunately, in many real-world applications, the true label distribution typically is imbalanced across classes involved in the collected data. Therefore, in this paper, we address the problem of learning from crowdsourced labeled data with an imbalanced true label distribution. We propose a new framework, named "Learning from Imbalanced Crowdsourced Labeled Data" (ICED), which simultaneously infers true labels from imbalanced crowdsourced labeled data and achieves high accuracy on downstream tasks such as classification. The ICED framework consists of two modules— a true label inference module and a synthetic data generation module— that augment each other iteratively. Extensive experiments conducted on both synthetic and real-world datasets demonstrate the effectiveness of the ICED framework. We will release datasets and code used for evaluation based on the acceptance of this paper.

Keywords: imbalanced data, crowdsourcing

1 Introduction

The success of supervised deep learning models in many real-world applications, such as image classification [17, 27, 28, 14] and speech recognition [8, 1, 12], is inseparable from the availability of large-scale labeled training data. However, obtaining a large amount of labeled data is often challenging. Annotating certain types of data samples such as medical images require specific domain knowledge [30], while some other types of data such as videos or audios are expensive in terms of time [33]. By inviting multiple crowd workers to annotate labels for data samples simultaneously or sequentially, modern crowdsourcing platforms, such as Amazon Mechanical Turk[1], offer a cost-effective way to collect large-scale labeled data [25]. Although crowdsourcing alleviates the label shortage problem to some extent, the annotated labels can be very inconsistent and noisy due to the cognitive differences between crowd workers [7]. For example, non-experts and experts may annotate the same object with distinct labels. As most existing supervised deep learning models only work well with determinate noise-free labels, there is a need for alternative approaches to handle such noisy labeled data.

In the past few decades, several methods have addressed noisy crowdsourced labels. One class of methods infer determinate true labels from crowdsourced labels [6, 24, 32]. Another class of methods learn a discriminative model directly from crowdsourced labeled data [25, 16, 29]. All the above approaches assume that the given training set is class-balanced, which is not true in real-world applications [31, 22], where *majority classes* have a significantly higher number of data samples than *minority classes*. Hence, those approaches perform poorly when training on imbalanced datasets.

There have been many attempts to address the challenges brought by imbalanced datasets, such as resampling approaches [21, 4, 10] and re-weighting approaches [5, 3]. However, these approaches require determinate noise-free training labels. Hence, they fail in crowdsourcing applications and there is a need for a new approach to address both imbalanced and noisy data. Based on that, in this paper, we study the problem of learning from imbalanced crowdsourced labeled data. To the best of our knowledge, this is the first work to learn an effective discriminative model on noisy labels when the latent true label distribution is imbalanced. Our goals are 1) to obtain accurate supervised information by inferring true labels from crowdsourced labels; 2) to ensure good prediction performance of the classifier on all classes in the balanced test set.

To meet these two goals, inspired by existing imbalanced data handling approaches and crowdsourced label processing approaches, we propose a novel framework ICED (Learning from Imbalanced Crowdsourced

*Michigan State University. {wangw116, xiaorui, tangjili}@msu.edu.

†University of Calgary. joseph.thekinen@ucalgary.ca.

‡TAL Education Group. liuzitao@tal.com.

[1] https://www.mturk.com

labEled Data). The ICED framework consists of two modules. One module uses generated synthetic data for minority classes to improve the true label inference process. Another module uses the inferred true labels to improve the quality of generated synthetic data. These two modules augment each other and improve themselves iteratively. After training, ICED is able to learn a classifier with good prediction performance on all classes uniformly distributed in the test set. The main contributions of this work are summarized below:

- We are the first one to address the problem of learning from imbalanced crowdsourced labeled data, a more realistic scenario in real-world.
- We present a novel framework ICED, which can simultaneously infer true labels from imbalanced crowdsourced labeled data and achieve good prediction performance on all classes.
- We conduct extensive experiments on both synthetic and real datasets to demonstrate the effectiveness of ICED on the classification task.

The rest of this paper is organized as follows. Sec. 2 presents our ICED framework in detail. Extensive empirical studies are introduced in Sec. 3. We summarize related works in Sec. 4 and conclude our work in Sec. 5.

2 The Proposed Framework

In this section, we first formulate the problem we studied and then introduce our proposed ICED framework.

2.1 Problem Formulation Suppose for a set of data samples $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, W workers are invited to annotate every sample in $\mathbf{X}$ and, hence, produce a set of crowdsourced labels $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, where $\mathbf{y}_i = \{(y_{i,1}, w_{i,1}), (y_{i,2}, w_{i,2}) \ldots, (y_{i,W}, w_{i,W})\}$. Here each annotation pair $(y_{i,u}, w_{i,u})$ represents label $y_{i,u}$ provided by worker w_u for sample $\mathbf{x}_i$ from C classes.

Definition 1 (**Learning from Imbalanced Crowdsourced Labeled Data**). Given a sample set $\mathbf{X}$ and its corresponding crowdsourced label set $\mathbf{Y}$, our goal is to obtain a classifier $\mathcal{F}$, which can achieve good prediction performance on uniformly distributed test data.

For each data sample $\mathbf{x}_i \in \mathbf{X}$, we assume it has W annotated labels. Moreover, as the true labels for sample set $\mathbf{X}$ is unknown, we denote the estimated true labels inferred by our ICED framework for $\mathbf{X}$ as $\mathbf{T} = \{t_1, t_2, \ldots, t_n\}$. In this paper, we focus on the binary classification task, i.e., there are two classes in the sample set $\mathbf{X}$ and, one is the majority class and the other is the minority class. Note that, our ICED framework is also adaptable for multi-class classification tasks with slight modifications and we will leave it as one future work.

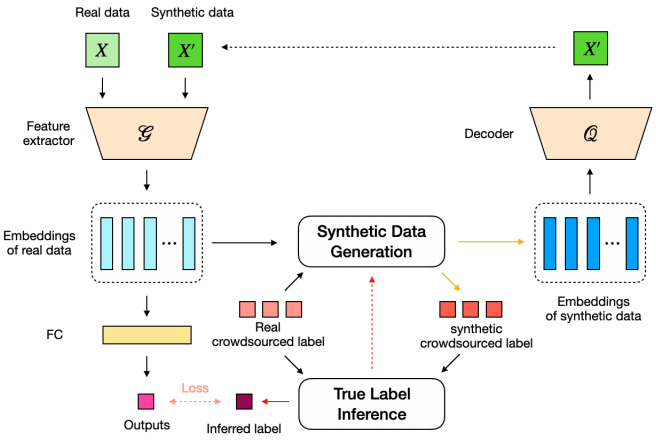


Figure 1: An overview of the ICED framework. The solid yellow and red arrows indicate outputs of the synthetic data generation module and the true label inference module, respectively, in the current training iteration. The red and black dash arrows represent the inferred labels and synthetic data samples, respectively, in the previous iteration.

2.2 Framework Overview For tackling the learning from imbalanced crowdsourced labeled data problem, we propose a novel framework ICED as shown in Figure 1. The main structure of ICED is a deep neural network based classifier $\mathcal{F}$ consisting of a feature extractor $\mathcal{G}$ and a fully connected layer (FC). During training $\mathcal{F}$, ICED introduces two modules: *true label inference* module and *synthetic data generation* module. The former estimates determinate true labels from given crowdsourced labeled data, and the latter generates synthetic data samples for the minority class using the estimated true labels. These two modules augment each other and improve themselves iteratively. Furthermore, to make the ICED framework obtain better initial learning ability at the beginning of the model training phase, ICED also includes a warm-up training strategy specifically designed for the crowdsourced labeled data. Next, we introduce details of each component.

2.3 True Label Inference Many classical approaches for inferring true labels from crowdsourced labels ignore the correlation between data samples and cognitive differences between individual crowd workers. For example, some workers tend to judge class c_α as class c_β by mistake due to their cognitive differences. Therefore, for overcoming the aforementioned short-

ages, our ICED framework adopts an EM approach [6] into the true label inference module to estimate determinate true labels from given crowdsourced labeled data.

To capture the annotation behaviors of crowd workers, we define $\Psi_{w_u}(c_\alpha, c_\beta)$ as the probability that worker w_u will annotate data samples with true label c_α as class c_β. Therefore, $\sum_{c_\beta \neq c_\alpha} \Psi_{w_u}(c_\alpha, c_\beta)$ represents the annotation error rate of the worker w_u when true label of samples are c_α. Let $\mathbb{T}$ be the random variable representing the true label of dataset $\mathbf{X}$ (similarly $\mathbb{T}_i$ for sample $\mathbf{x}_i$) and $\Phi_{c_\alpha} = p(\mathbb{T} = c_\alpha) = p(\mathbb{T}_i = c_\alpha)$ be the prior of class c_α, in the absence of any observations.

Our task is to estimate the probability of each label $c_\alpha \in [C]$ to be the latent true label for each sample $\mathbf{x}_i$ based on the crowdsourced labels $\mathbf{Y}$, i.e., $p(\mathbb{T}_i = c_\alpha|\mathbf{Y})$. The label with maximal probability is then chosen as the estimated label to train the classifier $\mathcal{F}$. The steps in the EM algorithm to estimate true labels are:

- E-step: computes the likelihood function of the observed crowdsourced labels $\mathbf{Y}$ based on current estimated true labels $\mathbf{T}$ and parameters $\Psi = \{\Psi_{w_u}(c_\alpha, c_\beta)|w_u \in [W], c_\alpha, c_\beta \in [C]\}$ and $\Phi = \{\Phi_{c_\alpha}|c_\alpha \in [C]\}$;
- M-step: updates the parameters by maximizing the likelihood function and refine the estimated true labels with new parameters.

In detail, we assume annotations provided by crowd workers are independently distributed. Given the current estimated true labels $\mathbf{T}$ and parameters Ψ and Φ, the likelihood of the observed crowdsourced labels $\mathbf{Y}$ can be obtained by

$$Q(\mathbf{Y}|\Psi, \Phi, \mathbf{T}) \propto \prod_{i\in[n]} p(\mathbb{T}_i = t_i) \prod_{u\in[W]} \Psi_{w_u}(t_i, y_{i,u}), \tag{2.1}$$

where i, u are the indices of data sample and crowd worker, respectively; $[n]$ and $[W]$ denote the sets $\{1, 2, \ldots, n\}$ and $\{1, 2, \ldots, W\}$, respectively.

The parameters in Ψ and Φ are updated by maximizing the above likelihood function. Specifically, Ψ can be computed as

$$\Psi_{w_u}(c_\alpha, c_\beta) = \frac{d(w_u, c_\alpha, c_\beta)}{d(w_u, c_\alpha)},$$

where $d(w_u, c_\alpha, c_\beta)$ represents the number of data samples labeled as c_β by worker w_u when their current estimated true labels is c_α, and $d(w_u, c_\alpha)$ represents the number of data samples labeled by worker w_u when their current estimated true labels is c_α. In addition, Φ can be computed as

$$\Phi_{c_\alpha} = \frac{\#\text{ samples whose true label is estimated as } c_\alpha}{\#\text{ samples in sample set } \mathbf{X}}.$$

Based on these updates, we can refine the estimation of true label by Bayes's theorem

$$\begin{aligned} p(\mathbb{T}_i = c_\alpha|\mathbf{Y}, \Psi, \Phi) &\propto p(\mathbf{Y}|\Psi, \Phi, \mathbb{T}_i = c_\alpha) p(\mathbb{T}_i = c_\alpha) \\ &\propto p(\mathbb{T}_i = c_\alpha) \prod_{u\in[W]} \Psi_{w_u}(c_\alpha, y_{i,u}), \end{aligned} \tag{2.2}$$

and choose the label c_α with largest probability as the current estimated true label for data sample $\mathbf{x}_i$. We repeat E-step and M-step iteratively until convergence.

In summary, the true label inference module can provide two important information for our ICED framework: 1) an estimation of latent true labels, $\mathbf{T}$, which can be used as supervised labels to train the classifier $\mathcal{F}$ and 2) the marginal distribution $p(\mathbb{T} = c_\alpha)$, which models the data imbalance between classes and thereby guiding the synthetic data generation module to augment a balanced synthetic dataset. Moreover, we also obtain the annotation error rate of each worker w_u, derived from $\Psi_{w_u}(\cdot, \cdot)$ as a by-product from the EM algorithm. The annotation error can be potentially used to penalize the unqualified workers whose error rate is relatively high, depending on application scenarios.

2.4 Synthetic Data Generation The performance of the EM approach adopted in the true label inference module depends on the choice of prior probability, e.g., Φ_{c_α}, for initialization. Conventionally, uniform prior is used for initialization, resulting in poor performance on imbalanced crowdsourced labeled training sets. Motivated by over-sampling approaches as an effective solution to handle imbalanced datasets, we integrate a synthetic data generation module in our ICED framework to balance the training set.

As shown in Figure 1, we first apply the true label inference module to obtain estimated true labels $\mathbf{T}$. We then use the feature extractor $\mathcal{G}$ in the deep neural network based classifier $\mathcal{F}$ to map all data samples in $\mathbf{X}$ from the raw data space into a latent embedding space. Finally, we apply the following synthetic data generation process in the embedding space.

Suppose $\mathbf{z}_i$ be the embedding of the data sample $\mathbf{x}_i$ in the learned latent space. Based on the information involved in the estimated true labels $\mathbf{T}$, all embeddings $\mathbf{z}_i$ belong to the minority class (determined by its corresponding estimated true label t_i) will be selected as candidate embeddings to help generate synthetic minority samples. After that, we use the linear interpolation operations adopted in the SMOTE [4] approach as a way to create synthetic minority sample embeddings. Specifically, for any candidate embedding $\mathbf{z}_i$, we (i) discover k nearest neighbors $\{\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^k\}$ for $\mathbf{z}_i$; and (ii) randomly pick up one nearest neighbor $\mathbf{z}_i^r$ from the set $\{\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^k\}$ to create a synthetic minority sample

embedding $\mathbf{z}'_i$ as follows:

$$\mathbf{z}'_i = \mathbf{z}_i + \delta\left(\mathbf{z}^r_i - \mathbf{z}_i\right), \tag{2.3}$$

where δ is a scalar in range $[0, 1]$. The step (ii) can repeat R times, and, finally, $R \times m$ synthetic minority sample embeddings will be generated when executing the same process on all selected candidate embeddings with size m.

Since the true label inference module cannot guarantee 100% accuracy on estimating latent true labels from crowdsourced labels, the estimated determinate labels still have a chance to be opposite to the latent true labels. Hence, to reduce the adverse effects of possible wrong inference, different from the SMOTE approach, which chooses δ randomly, we assign the value for δ based on the label certainty score of a data sample $\mathbf{x}_i$ and its selected neighbor $\mathbf{x}^r_i$.

Definition 2 (**Label Certainty Score**)**.** Given a data sample $\mathbf{x}_i$ and we assume that its crowdsourced label $\{y_{i,u} \mid u \in [W]\}$ follow the multinomial distribution. The label certainty score $S(\mathbf{x}_i)$ is defined as the inverse variance of this distribution and is computed as

$$S(\mathbf{x}_i) = \frac{1}{\mathbb{E}_u \|y_i - \mathbb{E}_u(y_i)\|^2 + \epsilon}, \tag{2.4}$$

where $\mathbb{E}_u$ is the expectation over crowdsourced label $\{y_{i,u} \mid u \in [W]\}$ for sample $\mathbf{x}_i$, and ϵ is a small constant to avoid numerical issue.

The label certainty score measures the agreement degree among crowd workers. Specifically, the label certainty score reaches its minimum value when a tie or a draw happens and goes to its maximum value when all annotated labels for one data sample are consistent.

Therefore, given sample $\mathbf{x}_i$ and its neighbor $\mathbf{x}^r_i$, the value for δ can be calculated by

$$\delta = S(\mathbf{x}_i) / \big(S(\mathbf{x}_i) + S(\mathbf{x}^r_i)\big) + \eta, \tag{2.5}$$

where η is sampled from a uniform distribution to add some randomness on the scalar δ. With the help of Eq. (2.5), the generated synthetic embeddings $\mathbf{z}'_i$ will be close to the candidate embedding, which has a higher label certainty score, such that we increase the probability of the generated embedding $\mathbf{z}'_i$ of being in the cluster of minority embeddings, and thereby alleviating imbalance issues.

After generating the synthetic minority sample embeddings, we apply the k-NN approach into the latent embedding space to construct synthetic crowdsourced labels for generated sample embeddings. More specifically, for any generated minority embedding $\mathbf{z}'_i$, we collect crowdsourced labels of its k nearest neighbor embeddings of real data samples. We then determine its synthetic crowdsourced labels by simulating the annotation behavior of each crowd worker in these collected k crowdsourced labels.

We obtain synthetic minority sample embeddings and corresponding synthetic crowdsourced labels, as shown in Figure 1. We then map the synthetic embeddings back to the raw data space using a pre-trained decoder $\mathcal{Q}$ and update the parameters of the classifier $\mathcal{F}$ using the augmented balanced training.

In summary, the synthetic data generation module in our ICED framework addresses the issues caused by the imbalanced training set by generating sufficient synthetic minority data samples and synthetic crowdsourced labels, benefitting both the true label inference process and the classifier training process.

2.5 Warm-up Training Strategy Recent studies have discovered that deep neural networks can learn even on noisy labeled data [23, 19]. Hence, a warm-up training phase is an effective strategy to initialize supervised deep learning models. Existing literature [18, 11] uses all available data in the warm-up training phase. Different from existing literature, in our ICED framework, we design a new warm-up training strategy for the crowdsourced labeled data.

Specifically, given sample set $\mathbf{X}$ and crowdsourced label set $\mathbf{Y}$, we first calculate the label certainty score for each data sample $\mathbf{x}_i$. After gathering label certainty scores for all data samples, we apply majority voting (MV) on the crowdsourced label set $\mathbf{Y}$ to obtain an estimated true label set $\mathbf{T}$. Each element t_i in $\mathbf{T}$ is obtained by aggregating its corresponding crowdsourced label $\mathbf{Y}_i$ using MV. We then divide the sample set $\mathbf{X}$ and corresponding true label set $\mathbf{T}$ into four different groups based on the label certainty scores: low certainty group, third-highest certainty group, second-highest certainty group, and highest certainty group. We use all data samples except those in the low certainty group to initially train the classifier $\mathcal{F}$ with associated determinate labels in a supervised way. The algorithm of our designed warm-up training strategy (Algorithm 2) can be found in Appendix A.

In general, there is a higher probability of the determinate label t_i, obtained by MV, being the same as the latent true label when the sample $\mathbf{x}_i$ has a higher label certainty score $S(\mathbf{x}_i)$. Our designed warm-up training strategy is similar to using noisy labeled data to help provide some initial abilities for the deep neural network based classifier $\mathcal{F}$ and using clean labeled data to fine-tune $\mathcal{F}$. After the warm-up training phase, our ICED framework is able to get a better initial learning abilities.

2.6 Algorithm In this subsection, we present the ICED framework in Algorithm 1.

Algorithm 1 The algorithm of ICED

Input: sample set $\mathbf{X}$, crowdsourced label set $\boldsymbol{Y}$

1: Conduct warm-up training. // Algorithm 2
2: **repeat**
3: Generate synthetic minority samples and corresponding crowdsourced labels. // Sec. 2.4
4: Obtain the inferred true label set $\mathbf{T}'$ for the augmented crowdsourced labels. // Sec. 2.3
5: Train the classifier $\mathcal{F}$ using the augmented data samples and inferred determinate labels $\mathbf{T}'$.
6: **until** model converge or maximum epoch reached

As shown in Algorithm 1, we first introduce our designed warm-up training strategy to make the deep neural network based classifier $\mathcal{F}$ obtain better initial abilities. Then, in each training epoch, we apply the synthetic data generation module, to produce synthetic minority samples with synthetic crowdsourced labels, for balancing the training set. After that, the true label inference module infers latent true labels for the augmented crowdsourced labels. Hence, the parameters of the classifier $\mathcal{F}$ can be updated based on the augmented balanced data samples and corresponding inferred determinate labels in a supervised way. We continuously conduct this iteration process until $\mathcal{F}$ converges or the maximum training epoch reaches.

3 Experiment

In this section, we conduct experiments to verify the effectiveness of our proposed ICED framework by answering the following three questions:

1. Can the proposed framework obtain good prediction performance on the balanced test data?
2. Does the generated synthetic data improve the accuracy of the true label inference process?
3. Does our newly designed warm-up training strategy improve over existing warm-up training strategies?

To answer the first question, we compare the performance of ICED with several state-of-the-art crowdsourced label processing approaches on the classification task. For the second question, we compare the accuracy of true label inference with and without synthetic data generation modules on two synthetic datasets. Finally, we compare the prediction performance of the deep neural network based classifier $\mathcal{F}$ using our designed warm-up training strategy and by traditional warm-up training strategies to answer the third question.

Table 1: Statistics of datasets. The entries in "# majority class" and "# minority class" represent the number of samples we used for those classes, respectively, to construct a synthetic training dataset.

Statistic item	Dataset			
	Gisette-Syn	USPS-Syn	GSAD-Syn	Emotion
# features	5,000	256	128	1,582
# training data	3,080	734	2,575	3,027
# majority class	2,800	668	2,341	-
# minority class	280	66	234	-
# crowd worker	7	9	11	5
# test data	1,400	332	1,170	900

3.1 Datasets

3.1.1 Synthetic Datasets We conduct experiments on three synthetic datasets and one real-world dataset. Table 1 summarizes key statistical information of these four datasets. The three synthetic imbalanced crowdsourced labeled datasets are constructed based on three widely used datasets: Gisette, USPS, and Gas Sensor Array Drift (GSAD). Specifically, Gisette and USPS datasets are from Feature Selection data repository[2] and the GSAD dataset is from UCI data repository[3]. Due to the limited space, we provide the details of constructing synthetic imbalanced crowdsourced labeled datasets used in our experiments in Appendix B.

3.1.2 Real Dataset We collected a real-world imbalanced crowdsourced labeled dataset *Emotion* from our educational practice. The collected data samples in the Emotion dataset are 1-minute audio tracks collected from multiple teachers who teach courses such as Mathematics and English in primary school. We split all audio tracks in Emotion into a training set and a test set with sample size 3,027 and 900 separately. Five teaching professionals are invited to annotate every audio track in the training set as either high emotion arousal or low emotion arousal to assess teaching effects on courses and the annotation results provided by one teaching expert for audio tracks in the test set are adopted as the ground truth labels. For experiment purpose, we maintained the same number of data samples in each class in the test set. More details about this real-world crowdsourced labeled dataset can be found in Appendix C.

3.2 Performance Comparison

3.2.1 Baseline Methods For evaluating the effectiveness of our proposed ICED framework on the learning from imbalanced crowdsourced labeled data prob-

[2] http://featureselection.asu.edu/datasets.php
[3] https://archive.ics.uci.edu/ml/index.php

Table 2: Classification performance of our ICED framework and baseline methods on four datasets.

	Gisette-Syn		USPS-Syn		GSAD-Syn		Emotion	
Methods	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
MV+LR	0.8179	0.8175	0.8494	0.8480	0.7094	0.6872	0.8289	0.8277
MV+DNN	0.8000	0.7979	0.8735	0.8732	0.8333	0.8327	0.7944	0.7901
D&S+LR	0.7671	0.7592	0.8193	0.8136	0.6974	0.6716	0.8311	0.8294
D&S+DNN	0.7636	0.7562	0.7530	0.7386	0.8085	0.8082	0.7811	0.7749
Crowd-Layer	0.8200	0.8167	0.9006	0.8998	0.8342	0.8295	0.8300	0.8273
MBEM	0.6967	0.4211	0.7813	0.5334	0.6826	0.5577	0.6344	0.5324
CPC	0.8021	0.8020	0.8313	0.8311	0.6154	0.5917	-	-
ICED	**0.8521**	**0.8512**	**0.9036**	**0.9030**	**0.8872**	**0.8865**	**0.8644**	**0.8640**

lem, we compare the performance of ICED with several representative state-of-the-art crowdsourced label processing approaches on the classification task, including:

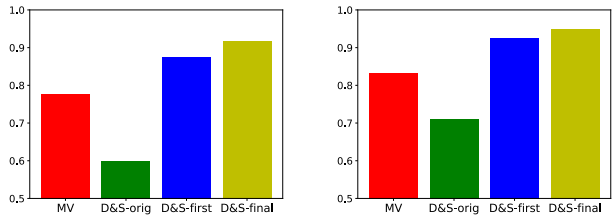
- Majority Voting (MV), which infers determinate labels based on the majority of annotated labels.
- D&S [6], which infers determinate labels via estimating the error rate of each crowd worker.
- Crowd-Layer [25], which is an end-to-end deep neural network containing a novel crowd layer to learn from crowdsourced labeled data directly.
- MBEM [16], which is able to learn from crowdsourced labeled data via jointly modeling latent true labels and crowd worker qualifications.
- CPC [15], which improves the performance of classifier via learning parameters of classifier and clusters of crowd workers jointly.

As MV and D&S can only infer determinate labels instead of learning a classifier from crowdsourced labels, we introduce two classifiers Logistic regression (LR) and deep neural networks (DNN). Specifically, we train LR and DNN on the same datasets with determinate labels inferred by MV and D&S individually and use them as baseline methods. We denote these baseline methods as *MV+LR*, *MV+DNN*, *D&S+LR* and *D&S+DNN*. The implementation details of aforementioned baseline methods as well as our proposed ICED framework are introduced in Appendix D.

Table 2 shows the classification performance of our ICED framework by comparing against seven baseline methods on three synthetic datasets and one real dataset. Based on this table, we have the following observations. Firstly, the classification performance of both LR and DNN, measured in terms of accuracy and F1-score, is higher when using MV instead of D&S to infer determinate labels. D&S, as an EM-based approach, assumes a uniform label distribution. MV independently aggregates annotated labels of each crowdsourced label. Hence, given an imbalanced crowdsourced labeled dataset, the performance of MV on the true label inference task will not be affected by the imbalanced true label distribution. On the contrary, D&S may show poor performance due to its inaccurate uniformity assumption. Secondly, our ICED framework achieves the best classification performance on all four datasets comparing with several representative state-of-the-art crowdsourced label processing approaches. We believe there are three reasons behind this performance. First, even though the D&S approach assumes uniformity in data distribution, ICED generates synthetic data to augment the imbalances between classes in the training set. The resulting training set will approximate a uniform distribution, enhancing the performance of the D&S approach. Second, the more accurate determinate labels inferred by the true label inference module improves the synthetic data generation module. The reason being, the synthetic data generation module can use the inferred determinate labels to differentiate minority data samples from majority ones to generate synthetic samples in minority classes. As a result, the data samples produced by the synthetic data generation module have a higher probability of belonging to the minority class. Third, the synthetic generated data can also help the classifier $\mathcal{F}$ in ICED to obtain better generalization ability during the model training phase via augmenting the imbalanced training set.

3.3 Ablation Study As we mentioned before, the D&S approach assumes uniform label distribution as prior knowledge for initialization. Therefore, the true label inference performance of the D&S approach is lower than the MV approach on the imbalanced crowdsourced labeled dataset. The ICED framework addresses the issue in the D&S approach by integrating a synthetic data generation module. The synthetic data generation module balances the imbalanced training set via generating synthetic data samples for minority classes. The resulting augmented dataset better fit the prior knowledge used in D&S.

To verify whether and how the synthetic generation



(a) Gisette-Syn. (b) GSAD-Syn.

Figure 2: Accuracy of true label inference using MV and the true label inference module (D&S) in ICED.

module benefits from the true label inference module in our ICED framework, we compare the true label inference accuracy of D&S adopted in ICED with MV. We show the comparison on two synthetic datasets— Gisette-Syn and GSAD-Syn— because the ground truth labels are available for these two datasets. In our experiments, we record the true label inference accuracy of D&S for three cases: 1) before introducing the synthetic data generation module, 2) after applying the synthetic data generation module once, and 3) after completing the training procedure of ICED. We denote these three cases as *D&S-orig*, *D&S-first*, *D&S-final*, respectively. In Figure 2, we find that the performance of D&S varies widely for different cases. Take the experimental results obtained on the training set of Gisette-Syn as an example. As shown in Figure 2(a), before introducing the synthetic data generation module, the true label inference accuracy of D&S is below 60%, which is much worse than MV. Surprisingly, by conducting the synthetic data generation process just once, the label inference accuracy of D&S is higher than 80%. After finishing the training procedure of ICED, i.e, after repeating the synthetic data generation process multiple times, D&S achieves higher than 90% true label inference accuracy, which is a significant improvement in comparison to a naive application of D&S on the imbalanced crowdsourced labeled dataset. In conclusion, the synthetic generation module significantly enhances the performance of the true label inference module in ICED.

3.4 Effectiveness of Warm-up Training In this subsection, we test the effectiveness of our designed warm-up training strategy. Given a set of crowdsourced labeled data, the warm-up training strategy adopted in our ICED framework first calculates label certainty score for each data sample based on its corresponding crowdsourced label. Then it divides data samples into different groups based on their label certainty scores. Data samples in the third-highest certainty group will feed the classifier $\mathcal{F}$ in ICED first with their corresponding determinate labels produced by MV. Data samples in the highest certainty group will train $\mathcal{F}$ after those in the second-highest group are picked. In experiments, we denote our designed warm-up training strategy as *ICED-w*. As a comparison, we implement one common warm-up training strategy used in literature for learning from noisy labeled data that uses all available data simultaneously to warm up the model. We denote this warm-up strategy as *Trad-I*. Another warm-up training strategy *Trad-II*, which is the same as Trad-I, except it only uses data samples in the highest, second-highest, and third-highest certainty groups rather than all the available data samples. In other words, Trad-II chooses the same data samples adopted in our designed warm-up training strategy ICED-w and uses them to feed $\mathcal{F}$ at the same time. For evaluation, we report the classification performance of $\mathcal{F}$ by training on different warm-up training strategies in Table 3. We observe that the classifier $\mathcal{F}$ training by ICED-w achieves the best classification performance, comparing with Trad-I and Trad-II, on all datasets. Thus, our designed warm-up training strategy more effectively initializes ICED.

Table 3: Performance of different warm-up strategies.

Datasets	Methods	# samples	# epoch	Accuracy	F1-score
Gisette-Syn	Trad-I	3,080	15	0.8014	0.7989
	Trad-II	2,043	15	0.6079	0.5372
	ICED-w	2,043	5×3	**0.8186**	**0.8126**
USPS-Syn	Trad-I	734	6	0.7500	0.7333
	Trad-II	103	6	0.7922	0.7828
	ICED-w	103	2×3	**0.8373**	**0.8329**
GSAD-Syn	Trad-I	2,575	6	0.4581	0.3142
	Trad-II	507	6	0.4504	0.3105
	ICED-w	507	2×3	**0.8376**	**0.8332**

4 Related Work

4.1 Processing Crowdsourced Label Inferring true labels from crowdsourced labels is a challenge as the crowd workers have diverse expertise [35]. A naive approach to infer true labels is majority voting (MV), which uses the majority of annotated labels as the true label. The MV approach performs poorly in practice, as the crowd workers have diverse expertise and reliability. An Expectation-Maximization (EM) [6] approach addresses the differences between crowd workers by estimating the error rate of each crowd worker from the crowd labels. Therefore, an EM approach has higher accuracy than MV in inferring true labels. Inspired by this, Whitehill et al. [32] used an iterative approach considering both sample difficulty and crowd worker reliability to infer true labels. The above approaches focus only on inferring true labels. Some recent works integrate true labels inference with downstream tasks. Kajino et al. [15] developed a clustered personal classifier method that simultaneously trains a classifier and esti-

mates a cluster of workers. Rodrigues et al. [26] generalized Gaussian process classification considering crowd workers with diverse expertise. Raykar et al. [25] designed an EM-based approach to jointly learn a crowd worker noise model and a regression model. Khetan et al. [16] proposed another EM-based approach for learning from crowdsourced labeled data by jointly modeling latent true labels and crowd worker qualification. Guan et al. [9] modeled information from each worker and then learned combination weights via back-propagation. As all the above approaches assume a uniformed label distribution as prior knowledge for initialization, they cannot achieve good generalization when the given training set has an imbalanced true label distribution.

4.2 Handling Imbalanced Data Existing approaches to handle imbalanced data mainly falls into two categories: re-sampling and re-weighting. Re-sampling approaches balance the imbalanced data through under-sampling data samples from majority classes [34, 21] or over-sampling data samples from minority classes [4, 10, 13]. As under-sampling approaches often discard several data samples, over-sampling approaches are better in practice. Synthetic Minority Over-sampling Technique (SMOTE) [4] is a well-accepted over-sampling approach. Instead of duplicating existing minority data samples to inflate minority classes, SMOTE produces unseen synthetic minority samples by applying linear interpolation operations between a specific minority sample and one of its nearest neighbors within the same class. Several variants of SMOTE [10, 13] further improve the prediction performance of classifiers training on imbalanced datasets. Re-weighting approaches allocate different weights for different classes or even different data samples. For example, Lin et al. [20] proposed Focal loss to reshape the standard cross entropy loss such that it down-weights the loss assigned to well-classified data samples. Cui et al. [5] presented to utilize the data overlap measurement to quantify the effective number of samples for each class and re-weight each class by the inverse of the number of effective samples per class. Existing imbalanced data handling approaches assume that the given labels are determinate and noise-free, which is not the case in crowdsourced scenarios. Therefore, learning from imbalanced crowdsourced labels needs to be addressed.

5 Conclusion

In this paper, we investigate the problem of learning from imbalanced crowdsourced labeled data. We present a novel ICED framework to deal with the imbalanced true label distribution and noisy crowdsourced labels. ICED framework alleviate the negative impacts of imbalanced true label distribution while using the supervised information in the crowdsourced labels. To evaluate the performance of the ICED framework, we apply ICED into a classification task by training on both synthetic and real imbalanced crowdsourced labeled datasets and compare its performance with several representative crowdsourced label processing approaches. Extensive experimental results demonstrate the effectiveness of our proposed framework ICED on learning from imbalanced crowdsourced labeled data.

In the future, we plan to modify our ICED framework to fit the multi-class imbalanced crowdsourced labeled data. In addition, we are also interested in exploring ICED based new framework to handle other imbalanced noisy labeled data, such as inexact labeled data.

References

[1] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, *Convolutional neural networks for speech recognition*, IEEE/ACM Transactions on audio, speech, and language processing, 22 (2014), pp. 1533–1545.

[2] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab, *Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images*, IEEE transactions on medical imaging, 35 (2016), pp. 1313–1321.

[3] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, *Learning imbalanced datasets with label-distribution-aware margin loss*, in Advances in Neural Information Processing Systems, 2019, pp. 1567–1578.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *Smote: synthetic minority over-sampling technique*, Journal of artificial intelligence research, 16 (2002), pp. 321–357.

[5] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, *Class-balanced loss based on effective number of samples*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.

[6] A. P. Dawid and A. M. Skene, *Maximum likelihood estimation of observer error-rates using the em algorithm*, Journal of the Royal Statistical Society: Series C (Applied Statistics), 28 (1979), pp. 20–28.

[7] J. Fan, G. Li, B. C. Ooi, K.-l. Tan, and J. Feng, *icrowd: An adaptive crowdsourcing framework*, in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015, pp. 1015–1030.

[8] A. Graves, A.-r. Mohamed, and G. Hinton, *Speech recognition with deep recurrent neural networks*, in 2013 IEEE international conference on acoustics, speech and signal processing, Ieee, 2013, pp. 6645–6649.

[9] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, *Who said what: Modeling individual labelers improves classification*, arXiv preprint arXiv:1703.08774, (2017).

[10] H. Han, W.-Y. Wang, and B.-H. Mao, *Borderline-smote: a new over-sampling method in imbalanced data sets learning*, in International conference on intelligent computing, Springer, 2005, pp. 878–887.

[11] J. Han, P. Luo, and X. Wang, *Deep self-learning from noisy labels*, in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5138–5147.

[12] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al., *Deep speech: Scaling up end-to-end speech recognition*, arXiv preprint arXiv:1412.5567, (2014).

[13] H. He, Y. Bai, E. A. Garcia, and S. Li, *Adasyn: Adaptive synthetic sampling approach for imbalanced learning*, in 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, 2008, pp. 1322–1328.

[14] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[15] H. Kajino, Y. Tsuboi, and H. Kashima, *Clustering crowds*, in Proceedings of the twenty-seventh AAAI conference on artificial intelligence, 2013, pp. 1120–1127.

[16] A. Khetan, Z. C. Lipton, and A. Anandkumar, *Learning from noisy singly-labeled data*, arXiv preprint arXiv:1712.04577, (2017).

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in Advances in neural information processing systems, 2012, pp. 1097–1105.

[18] K.-H. Lee, X. He, L. Zhang, and L. Yang, *Cleannet: Transfer learning for scalable image classifier training with label noise*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5447–5456.

[19] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, *Learning to learn from noisy labeled data*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5051–5059.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection*, in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[21] X.-Y. Liu, J. Wu, and Z.-H. Zhou, *Exploratory undersampling for class-imbalance learning*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39 (2008), pp. 539–550.

[22] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, *Large-scale long-tailed recognition in an open world*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2537–2546.

[23] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, *Learning with noisy labels*, in Advances in neural information processing systems, 2013, pp. 1196–1204.

[24] 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 (2010).

[25] F. Rodrigues and F. Pereira, *Deep learning from crowds*, arXiv preprint arXiv:1709.01779, (2017).

[26] F. Rodrigues, F. Pereira, and B. Ribeiro, *Gaussian process classification and active learning with multiple annotators*, in International conference on machine learning, 2014, pp. 433–441.

[27] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, (2014).

[28] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[29] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. Silberman, *Learning from noisy labels by regularized estimation of annotator confusion*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 11244–11253.

[30] T. Tommasi, F. Orabona, and B. Caputo, *Discriminative cue integration for medical image annotation*, Pattern Recognition Letters, 29 (2008), pp. 1996–2002.

[31] G. Van Horn and P. Perona, *The devil is in the tails: Fine-grained classification in the wild*, arXiv preprint arXiv:1709.01450, (2017).

[32] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, *Whose vote should count more: Optimal integration of labels from labelers of unknown expertise*, in Advances in neural information processing systems, 2009, pp. 2035–2043.

[33] G. Xu, W. Ding, J. Tang, S. Yang, G. Y. Huang, and Z. Liu, *Learning effective embeddings from crowdsourced labels: An educational case study*, in 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 1922–1927.

[34] S.-J. Yen and Y.-S. Lee, *Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset*, in Intelligent Control and Automation, Springer, 2006, pp. 731–740.

[35] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, *Truth inference in crowdsourcing: Is the problem solved?*, Proceedings of the VLDB Endowment, 10 (2017), pp. 541–552.