Learning Only When It Matters: Cost-Aware Long-Tailed Classification

Yu-Cheng He[1], Yao-Xiang Ding[2], Han-Jia Ye[1,3], Zhi-Hua Zhou[1,3]

[1]National Key Laboratory for Novel Software Technology, Nanjing University, China
[2]State Key Laboratory of CAD & CG, Zhejiang University, China
[3]School of Artificial Intelligence, Nanjing University, China
{heyc,yehj,zhouzh}@lamda.nju.edu.cn, dingyx.gm@gmail.com

Abstract

Most current long-tailed classification approaches assume the *cost-agnostic* scenario, where the training distribution of classes is long-tailed while the testing distribution of classes is balanced. Meanwhile, the misclassification costs of all instances are the same. On the other hand, in many real-world applications, it is more proper to assume that the training and testing distributions of classes are the same, while the misclassification cost of tail-class instances is varied. In this work, we model such a scenario as *cost-aware* long-tailed classification, in which the identification of high-cost tail instances and focusing learning on them thereafter is essential. In consequence, we propose the learning strategy of augmenting new instances based on adaptive region partition in the feature space. We conduct theoretical analysis to show that under the assumption that the feature-space distance and the misclassification cost are correlated, the identification of high-cost tail instances can be realized by building region partitions with a low variance of risk within each region. The resulting AugARP approach could significantly outperform baseline approaches on both benchmark datasets and real-world product sales datasets.

Introduction

Long-tailed classification, in which the training distribution of classes is highly imbalanced and long-tailed, has received great attention in recent years (Liu et al. 2019; Wei and Li 2018; Zhang et al. 2021; Buda, Maki, and Mazurowski 2018). The major challenge for long-tailed classification is the lack of data for the tail classes, making it difficult to obtain generalizable information from them. Meanwhile, this issue becomes significant when the performance on the tail data is essential. To model this situation, most current long-tailed classification studies assume that the testing distribution is class-balanced: the numbers of instances for all classes are the same. As a result, the performance under tail classes is focused on by increasing the ratio of tail instances in the testing dataset. Meanwhile, this setting is *cost-agnostic*: The misclassification cost of any instance is the same.

However, we argue that this is not the only natural scenario of long-tailed classification. In fact, class-imbalanced learning is a long-standing research topic in machine learning (Zhou 2021). In tradition, class-imbalanced learning usually appears in companion with cost-sensitive learning, (Elkan 2001; Zhou and Liu 2010) where the misclassification of minority class instances is higher than the majority class instances. This situation appears in many real-world applications such as object detection (Oksuz et al. 2020) and medical diagnosis (Mazurowski et al. 2008). This relationship is not incident: the minority instances are essential only when misclassifying them leads to higher costs. In consequence, we follow this tradition to study *cost-aware* long-tailed classification, where the training and testing class distributions remain the same, while the misclassification cost of instances is varied. Furthermore, we assume that not all tail class instances are essentially important: *only some of the tail class instances incur high costs*. This situation appears in many real-world applications. For example, for a classification model on an e-commerce platform, the costs of misclassifying different goods, which can be the sale prices, are varied even on the tail instances with low sale volumes. Some low-sale goods may also be low-priced so the misclassification cost is also low. On the other hand, what matters are indeed those low-sale high-cost goods, which should be the true targets to pay attention to. Furthermore, it is common in real-world applications that *the instance features and their cost are highly correlated*. For instance, in the e-commerce example, the high-price goods are likely to share common properties, making their features similar. Designing cost-aware learning algorithms by utilizing this phenomenon is also a meaningful challenge.

In this paper, we conduct both theoretical and algorithmic studies on cost-aware long-tailed classification. We focus on the central challenge of identifying the high-cost instances and focusing the learning process on them thereafter. To supply generalizable information on the high-cost parts of the tail distribution, we propose the learning strategy of augmenting new instances based on region partitions in the feature space, which are built online during the learning process. We first conduct the access risk analysis of the learned model from the region partition. Based on the assumption that the feature-space distance among instances and their misclassification costs are correlated, the risk bound indicates the region partition rule of minimizing the variance of risk in each region. Based on this finding, we propose the AugARP approach for adaptive region partition by introducing an attention module into the convolutional network, which can be used in companion with any current instance-augmentation approaches. We verify the effectiveness of AugARP under benchmark datasets, showing its effectiveness over existing cost-agnostic baselines. We further verify its potential usefulness in real-world applications on the Amazon Products Sales 2023 dataset.

Related Work

Long-Tailed Classification

Class-imbalanced learning is often tackled by *rescaling* (Zhou 2021), which can be realized by *reweighting* which assigns higher loss weighs to minority class examples, or *resampling* which over-samples minority class examples or down-samples majority class examples. There are many successful approaches such as SMOTE (Chawla et al. 2002) and EasyEnsemble (Liu, Wu, and Zhou 2009). However, it has been observed that simple re-weighting and re-sampling methods for deep neural networks lead to an unsatisfying performance in long-tailed classification tasks when using deep neural network classifiers (Cao et al. 2019; Kang et al. 2020; Zhou et al. 2020). One of the major discoveries is that simple re-weighting and re-sampling lead to significant over-fitting of the feature backbone to minority instances during training. To address this issue, Decoupling the training of feature backbones with prediction heads becomes a popular learning strategy (Kang et al. 2020). Besides the decoupled training strategy, several interesting techniques for long-tailed classification have been proposed in recent years. To name a few, some works propose to directly manipulate the decision boundaries for fixing the bias from class-imbalance (Ye et al. 2020; Kim and Kim 2020). While some studies proposed specific designs of loss functions for better re-weighting instances (Cui et al. 2019; Li et al. 2020; Tan et al. 2020). Margin-based approaches are also proposed for alleviating the over-fitting from aggressive re-weighting and re-sampling (Cao et al. 2019).

In spirit, the most related long-tailed classification approaches to our work are Remix (Chou et al. 2020) and MFW (Ye, Zhan, and Chao 2021), which make use of Mixup-style (Zhang et al. 2018) strategy to augment the training instances. The idea of synthesizing instances to address class imbalance can be traced back to SMOTE (Chawla et al. 2002). The principles behind MFW and Remix are different: MFW focuses on weakening the features of head classes, while Remix focuses on the augmentation of tail class instances by borrowing head class features. Both methods can serve as a sub-routine of our approach for augmenting instances, while neither of them is designed for the cost-aware scenario, where utilizing the cost information is essential for learning.

Cost-Sensitive Learning

Cost-sensitive learning is a machine learning branch that handles unequal costs occurring in the learning process. In addition to other costs, the most studied one is class-wise unequal misclassification cost (Elkan 2001; Liu and Zhou 2006), i.e., misclassifying different class examples may suffer from unequal costs. There are many tasks involving both class-imbalance and unequal classification costs, it is well known from Liu and Zhou (2006) that if there is only mild class-imbalance, no specific processing is needed; while if there is severe class-imbalance, *rescaling* can be helpful. Generally, the rescaling is executed by reweighting or resampling according to the sizes of all classes, assuming that all minority classes have higher costs. In this paper, we study an even more complicated task, where the classes are long-tailed, and only a part of the minority classes are with high costs.

Region-Based Active Learning

Our AugARP approach works by building online partitions of the region space and actively choosing promising regions for instance augmentation. The ideas of region partition and active learning are also considered in (Cortes et al. 2019, 2020). Cortes et al. (2019) introduced the divide-and-conquer strategy into the active learning area and proposed a region-based active learning algorithm named ORIWAL, which is based on the idea of querying samples on regions that are hard for classifiers to learn. Based on ORIWAL, Cortes et al. (2020) proposed ARBAL which adaptively divides the regions of ORIWAL. Due to the significant difference in learning setup, we present a very different approach.

Algorithmic Framework

In this section, we first introduce the problem setup of cost-aware long-tailed classification. Subsequently, we propose the algorithmic framework of our approach based on the decomposition of the cost-aware access risk.

Problem Setup

In cost-aware long-tailed classification, the learner is given a cost-aware labeled dataset $D = \{z_i = (\mathbf{x}_i, y_i, c_i)\}_{i=1}^N$, in which each instance $z = (\mathbf{x}, y, c)$ is a triplet of input x, ground-truth label y, and bounded mis-classification cost c. We assume that each instance is generated from the following process. The ground-truth label y is first generated from $\mathcal{D}(y)$, a long-tailed distribution of classes over the label space $\mathcal{Y}$. This assumption indicates that the numbers of instances in each class are highly imbalanced, such that there are a few numbers of *head* classes with sufficient training instances and a large number of *tail* classes with insufficient training instances. We denote the cardinality of $\mathcal{Y}$ by L. Given y, x and c are generated from two class-conditional distributions $\mathcal{D}(\mathbf{x}|y)$ and $\mathcal{D}(c|y)$. We don't assume $\mathcal{D}(c|y)$ to be any specific distribution, while it is natural to assume that the misclassification cost is generally similar within a class. In general, for one instance, we assume that x and c are generated independently when y is given.

During testing, we assume that the testing instances are generated from the same procedure as the training instances under $\mathcal{D}(y)$, $\mathcal{D}(\mathbf{x}|y)$, and $\mathcal{D}(c|y)$. Note that this is different from the commonly studied cost-agnostic setting, where sampling the testing instances from the uniform distribution of labels is necessary for emphasizing the performance of tail classes. While in our cost-aware setting, this is unnecessary since the cost is explicitly considered.

For learning under the above setting, we assume to use a set of classifiers $h \in \mathcal{H}$. Given an input $\mathbf{x}$, a classifier $h : \mathcal{X} \to \mathbb{R}^L$ outputs L scores $h(\mathbf{x}, y) \in [0, 1], y \in \mathcal{Y}$ over classes.

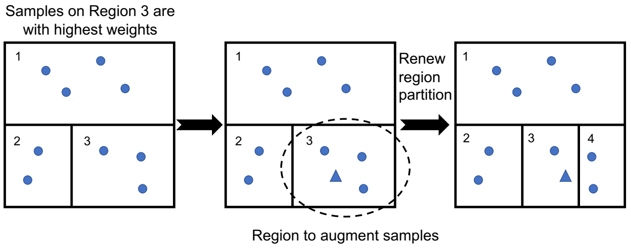

Samples on Region 3 are with highest weights
1
2
3
1
2
3
Renew region partition
1
2
3
4
Region to augment samples

Figure 1: The basic strategy of our proposed approach. The square space represents the feature space. The circles represent the feature embeddings of the original instances in the training dataset. The triangles represent new instances generated by augmentation.

The prediction is then given by $h(\mathbf{x}) = \arg\max_{y\in\mathcal{Y}} h(\mathbf{x}, y)$. Denote by $\bar{h}(\mathbf{x})$ the vector of $h(\mathbf{x}, y)$. We further assume that one classifier h can be represented by $\bar{h}(\mathbf{x}) = \sigma[W^T\phi(\mathbf{x})]$, in which $\phi : \mathcal{X} \to \mathbb{R}^d$ is the feature backbone outputting d-dimensional feature vectors, W is $d \times L$ matrix representing L d-dimensional linear prediction heads $[w_1; w_2; \ldots; w_L]$, and $\sigma : \mathbb{R}^L \to [0, 1]^L$ is a normalization function to ensure $h(\mathbf{x}, y)$ lying in $[0, 1]$. This formulation is compatible with most of the current neural network classifiers.

Region-Based Risk Decomposition

For one classifier h, We defined the *cost-aware classification error* of mis-classifying an instance $z = (x, y, c)$ as

$$\mathrm{CErr}(z; h) = c\mathbb{I}[h(\mathbf{x}) \neq y].$$

Denote $\mathcal{D}(z)$ as the instance generation distribution such that $\mathcal{D}(z) = \mathcal{D}(y)\mathcal{D}(\mathbf{x}|y)\mathcal{D}(c|y)$. We can define the empirical and expected *cost-aware classification risk* of classifier h over D and $\mathcal{D}(z)$:

$$\widehat{R}(h; D) = \frac{1}{N}\sum_{i=1}^{N}[\mathrm{CErr}(z_i; h)],$$
$$R(h; \mathcal{D}) = \mathbb{E}_{z\sim\mathcal{D}(z)}[\mathrm{CErr}(z; h)].$$

Bounding $R(h)$ by minimizing $\widehat{R}(h)$ is the golden-standard strategy in machine learning. While similar to common long-tailed classification, the major issue is that the number of instances from high-cost tail classes is small, which should be focused on during the learning process. On the other hand, under the cost-aware setting, there is an additional challenging task for distinguishing the high-cost tail classes from the low-cost ones. This can be accomplished by leveraging the correlation assumption between feature and cost: the instances with similar costs are relatively closer in the feature space. In general, our approach is based on the following basic strategy: *We can partition the feature space into disjoint regions. Based on the region partitions, better cost-aware performance can be achieved by augmenting new instances in the feature space within these high-cost regions,* as is shown in Fig. 1. For the construction of h, the augmentation is performed in the d-dimensional output space of $\phi(x)$. Notice that our proposed strategy relies on the correlation assumption between feature and cost, thus it is necessary to examine whether this assumption is consistent with reality. Due to the lack of space, we put the discussion on this issue in the appendix.

The central challenge for doing region-based feature augmentation is to partition the feature space properly. We conduct the following analysis to obtain theoretical insights on this task. Denote by $\mathcal{X}_\phi$ the feature space induced by $\mathcal{X}$ and ϕ. Furthermore, Denote by $\mathcal{D}_\phi(z)$ the feature space distribution over $\mathcal{X}_\phi$ induced by $\mathcal{D}(z)$ and ϕ. Assume that the feature space $\mathcal{X}_\phi$ is partitioned into K disjoint regions $R_1, R_2, ..., R_K$. We denote $p_1, p_2, \ldots, p_K$ as the probability ratio of $\mathcal{D}_\phi(z)$ over $R_1, R_2, \ldots, R_K$. We further denote $\{\mathcal{D}_{\phi,1}(z), \ldots, \mathcal{D}_{\phi,K}(z)\}, \{\mathcal{D}_1(z), \ldots, \mathcal{D}_K(z)\}$ as marginal feature and data distributions induced by $p_1, p_2, \ldots, p_K$. To proceed, given margin threshold $0 < \rho < 1$ and sub-samples D_k of size $N_k > 1$ such that each D_k is the subset of the whole dataset D in the k-th region, we introduce the *cost-aware margin error*

$$\mathrm{CErr}_\rho(z; h) = c\big[\Phi_\rho[\rho_h(x, y)]\big]$$

and *cost-aware margin risks*:

$$R_\rho(h; \mathcal{D}_k) = \mathbb{E}_{z\sim\mathcal{D}_k(z)}[\mathrm{CErr}_\rho(z; h)],$$
$$\widehat{R}_\rho(h; D_k) = \frac{1}{N_k}\sum_{i=1}^{N_k}[\mathrm{CErr}_\rho(z_i; h)],$$

in which $\rho_h(\mathbf{x}, y) = h(\mathbf{x}, y) - \max_{y'\neq y}[h(\mathbf{x}, y')]$ and

$$\Phi_\rho[a] = \begin{cases} 1, & a \leq 0, \\ 1 - a/\rho, & 0 < a < \rho, \\ 0, & a \geq \rho. \end{cases}$$

The following result shows that the region-decomposed access risk of the learned classifier:

Theorem 1. *Without loss of generality, assume that the costs c are bounded in $[0, 1]$, the normalization function σ is L_σ-Lipschitz, and the 2-norm of any $w \in W$ is bounded: $\|w\| \leq \Lambda$. Furthermore, denote by $h^* = \arg\min_{h\in\mathcal{H}} R(h; \mathcal{D})$, and assume that the learning algorithm can output a single $\widehat{h}$ such that $\exists 0 < \rho < 1, \forall k \in [K]$ (1) $\widehat{h} \in \arg\min_{h\in\mathcal{H}} \widehat{R}_\rho(h; D_k)$; (2) $\exists h_k^*, R_\rho(h_k^*; \mathcal{D}_k) = 0$. Then the following result holds with probability at least $1 - \delta$:*

$$R(\widehat{h}; \mathcal{D}) \leq R(h^*; \mathcal{D}) + 2\sum_{k=1}^{K} p_k \mathrm{Conf}_k,$$

in which

$$\mathrm{Conf}_k = \left(\frac{8LL_\sigma\Lambda}{\rho}\right)\sqrt{\frac{\mathrm{Var}_k(\widehat{\phi})}{N_k}} + \sqrt{\frac{2\,\mathrm{Var}_k(\widehat{h})\log\frac{3K}{\delta}}{N_k}} + \frac{\log\frac{3K}{\delta}}{3N_k}, \tag{1}$$

$\mathrm{Var}_k(\widehat{\phi}) = \mathbb{E}\big[\|\widehat{\phi}(x_i) - \mathbb{E}[\widehat{\phi}(x)]\|^2\big]$ *given that $\widehat{\phi}$ is the feature backbone of $\widehat{h}$. Furthermore,* $\mathrm{Var}_k(\widehat{h})$ *is the variance of cost-aware margin error under the k-th region:*

$$\mathrm{Var}_k(\widehat{h}) = \mathbb{E}\big[(\mathrm{CErr}_\rho(z; \widehat{h}) - \mathbb{E}[\mathrm{CErr}_\rho(z; \widehat{h})])^2\big].$$

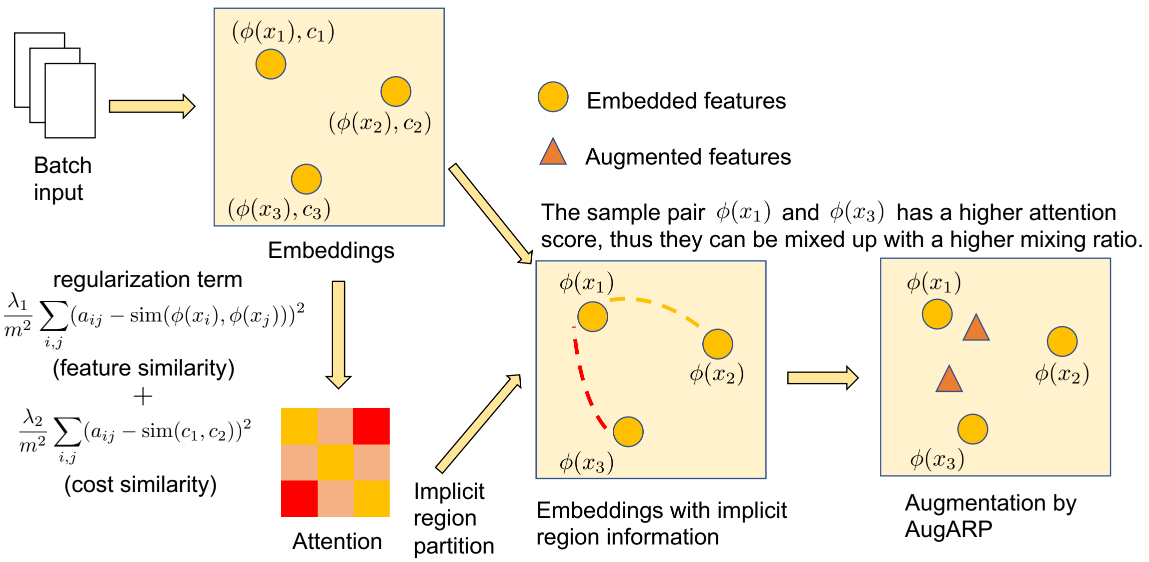

Batch input
$(\phi(x_1), c_1)$
$(\phi(x_2), c_2)$
$(\phi(x_3), c_3)$
Embeddings
Embedded features
Augmented features
The sample pair $\phi(x_1)$ and $\phi(x_3)$ has a higher attention score, thus they can be mixed up with a higher mixing ratio.
regularization term
$\frac{\lambda_1}{m^2}\sum_{i,j}(a_{ij} - \mathrm{sim}(\phi(x_i), \phi(x_j)))^2$
(feature similarity)
$+$
$\frac{\lambda_2}{m^2}\sum_{i,j}(a_{ij} - \mathrm{sim}(c_1, c_2))^2$
(cost similarity)
Attention
Implicit region partition
$\phi(x_1)$
$\phi(x_2)$
$\phi(x_3)$
Embeddings with implicit region information
$\phi(x_1)$
$\phi(x_2)$
$\phi(x_3)$
Augmentation by AugARP

Figure 2: Illustration of the proposed approach. An individual attention module is added after the feature-extracting modules. The attention module outputs the implicit region partition information, which is used to generate augmented samples. a_{ij} represents attention score, m represents batch size, and sim represents similarity measure.

Remark 1. The first message from Theorem 1 is that the excess risk of the learned model $\widehat{h}$ can be controlled by minimizing the empirical cost-aware margin risks under all the regions. Due to the power of current learning models, such as deep neural works, this target can be easily achieved. We also note that the cost-aware margin risk can be minimized even when a cost-free loss is used: The cost-aware loss gets to zero as long as the margin of the true label exceeds the threshold ρ.

Remark 2. The second message is the dependence of sample complexity over the variances of feature embeddings and risk values under each region. It can be observed that when both these variances are small in one region, then the order of sample complexity achieves the fast rate of $\mathcal{O}(1/N_k)$ instead of the ordinary $\mathcal{O}(1/\sqrt{N_k})$. Since our basic learning strategy is to augment instances, achieving this benefit is indeed important. This observation leads to the basic rule of region partition: the regions should be partitioned to minimize the feature and risk variances within each region. However, in general, finding a regime to minimize both variances is quite difficult. When feature and cost distributions have a close relationship, as assumed in our setting, we can achieve both targets by focusing on the risk variance only.

Augmentation by Adaptive Region Partition

In this section, based on the insight obtained from Section , we propose the AugARP (**Aug**mentation by **A**daptive **R**egion **P**artition) approach for cost-aware long-tailed classification. Algorithm 1 illustrates the running process of AugARP, which is based on two subroutines: `AUGMENT` and `PARTITION`. We assume that before running AugARP, the feature backbone of the classifier is learned from the original training dataset, as suggested by previous studies (Cao et al. 2019; Kang et al. 2020; Zhou et al. 2020). In each learning iteration, the training dataset is augmented by `AUGMENT`. After updating the prediction heads of the classifier, `PARTITION` proceeds to update the region partition. Below we discuss these two subroutines in detail.

Algorithm 1 AugARP

```
Input: region partition $\mathfrak{R} = \{\mathcal{X}_{\widehat{\phi}}\}$, augmentation counts $M$, embedded training dataset $D_{\widehat{\phi}}$, classifier $\widehat{h}$ with learned feature backbone $\widehat{\phi}$;
for step $t \in [T]$ do
    Instance augmentation:
        $(D^{aug}_{\widehat{\phi}}, \mathcal{X}_{\widehat{\phi},k^*}) \leftarrow \texttt{AUGMENT}(\mathfrak{R}, D_{\widehat{\phi}}, \widehat{h})$;
    Dataset update: $D_{\widehat{\phi}} \leftarrow D_{\widehat{\phi}} \cup D^{aug}_{\widehat{\phi}}$;
    Model update: Update prediction heads $\widehat{W}$ of $\widehat{h}$ using $D_{\widehat{\phi}}$;
    Region partition: $\mathfrak{R} \leftarrow \texttt{PARTITION}(\mathfrak{R}, D_{\widehat{\phi}}, \mathcal{X}_{\widehat{\phi},k^*}, \widehat{h})$;
end for
Output: Learned classifier $\widehat{h}$;
```

Adaptive Region Partition

The naive region partition strategy is to traverse the possible partition points for each feature, and then choose the partition feature and point with the highest reduction of cost variance.

However, the naive strategy described above is very inefficient, especially on large datasets with a huge feature space. In order to achieve efficient partitioning, we propose an adaptive region partition method based on the attention mechanism. As shown in Fig. 2, we add an attention module after all convolution modules in CNN. The function of the

Algorithm 2 AUGMENT

```
Input: attention module Att, embedded training dataset $D_{\widehat{\phi}}$, classifier $\widehat{h}$, beta distribution coefficient $\alpha$;
while not converge do
    Permute $D_{\widehat{\phi}}$ to generate $D'_{\widehat{\phi}}$;
    for $n = 1, 2, ..., |D_{\widehat{\phi}}|$ do
        $\lambda_n \sim \mathrm{Beta}(\alpha, \alpha)$;
        $\lambda_n = 2 \cdot \mathrm{sigmoid}(\texttt{Att}(\mathbf{x}_n, \mathbf{x}'_n))\lambda_n \cdot s(N_{y_n})$;
        $\mathbf{x}_n^{aug} = (1 - \lambda_n)\mathbf{x}_n + \lambda_n \mathbf{x}'_n$;
        $y_n^{aug} = y_n$;
    end for
end while
Return: $D^{aug}_{\widehat{\phi}}$, $\mathcal{X}_{\widehat{\phi},k^*}$;
```

attention module is to implicitly determine the region partition by calculating the similarity between two samples, as a high attention score of the two samples indicates that the two samples are in the same region. The attention module is trained by the similarity of sample features and costs to fit the implicit region information.

Region-Based Augmentation

In this subsection, we design the augmentation subroutine by combining the implicit region partition information with Mixup augmentation algorithms. Algorithm 2 provides an example of `AUGMENT` procedure in which we introduce the implicit region partition information into the MFW algorithm (Ye, Zhan, and Chao 2021) and increase the mixup proportion for samples with high attention scores, as the high attention scores indicates that these samples are in the same region. The definition of $s(N_{y_n})$ is the same as the definition in Ye, Zhan, and Chao (2021).

Besides MFW, other Mixup algorithms can also be combined with our proposed AugARP framework. In the experiment part, we investigate the performance of AugARP combined with MFW and Remix (Chou et al. 2020).

Experiment

In this section, we verify our proposed AugARP in the following three perspectives: (1) comparing the performance between AugARP and other long-tail learning methods; (2) whether the high-cost samples can be separated by `PARTITION`; (3) the effectiveness of proposed `AUGMENT` method. Four datasets are used in our experiments: CIFAR10, CIFAR100, iNaturalist, and 2023 Amazon Sales Dataset[1]. We introduce briefly each dataset and experiment setting in their respective subsections[2].

Performance Measure

In our experiments, the performance of each algorithm is measured by the weighted accuracy of the test set, which is the normalized version of CErr. Denote the test set as $(\mathbf{x}_i, y_i, c_i)_{i=1}^{N}$, in which $\mathbf{x}_i$ represents features, y_i represents the ground-truth label, and c_i represents the cost. The weighted accuracy of classifier h is defined as

$$\begin{aligned}\mathrm{WeightedAcc}(h) &= \frac{\sum_{i=1}^{N} c_i \mathbb{I}(h(\mathbf{x}_i) = y_i)}{\sum_{i=1}^{N} c_i} \\ &= 1 - \frac{\hat{R}(h)}{\frac{1}{N}\sum_{i=1}^{N} c_i}\end{aligned}$$

The proportion of samples belonging to each class in the test dataset is the same as the training set. Notice that the samples belonging to each class are not the same, while the proportion of each class in the test set is equaled in other related works. The reason is that the setting of our work is different from others, as we have assigned misclassification costs for samples in the dataset.

Comparison Methods

We compared our method with both the re-sampling-based methods and the re-weighting-based methods. The re-sampling-based methods include MFW (Ye, Zhan, and Chao 2021) and Remix (Chou et al. 2020). As our proposed AugARP algorithm also combines with MFW and Remix, the experiment verified the effect of our proposed region-based framework, and whether the region-based framework can be combined with MFW and Remix. The re-weighting-based method includes the basic ERM method with the cross-entropy loss, Focal (Lin et al. 2020), LDAM, and LDAM-DRW (Cao et al. 2019).

Experiment Results on Long-Tailed CIFAR10/CIFAR100 Datasets

This subsection introduces the experiment results on imbalanced CIFAR10/CIFAR100 datasets. CIFAR10 and CIFAR100 datasets consist of 32 × 32 images, 50000 images in the training set, and 10000 images in the test set. We follow (Cao et al. 2019) to construct the long-tail setting and step setting imbalanced CIFAR10/CIFAR100 training datasets, and the imbalance ratio is set as 100 for both the long-tail setting and the step setting. On CIFAR100, 10 head classes and 40 tail classes are set as important classes with misclassification costs of 100 while others are with misclassification costs of 1. The misclassification costs of CIFAR10 is similar to CIFAR100. Unlike other papers that use a balanced test set, we have set the weight of each class, so we set the sample proportion of the test set to be the same as the training set. We use Resnet-32 for both CIFAR10 and CIFAR100 datasets. For each algorithm, we train 200 epochs and repeat 3 times to report the average weighted accuracy.

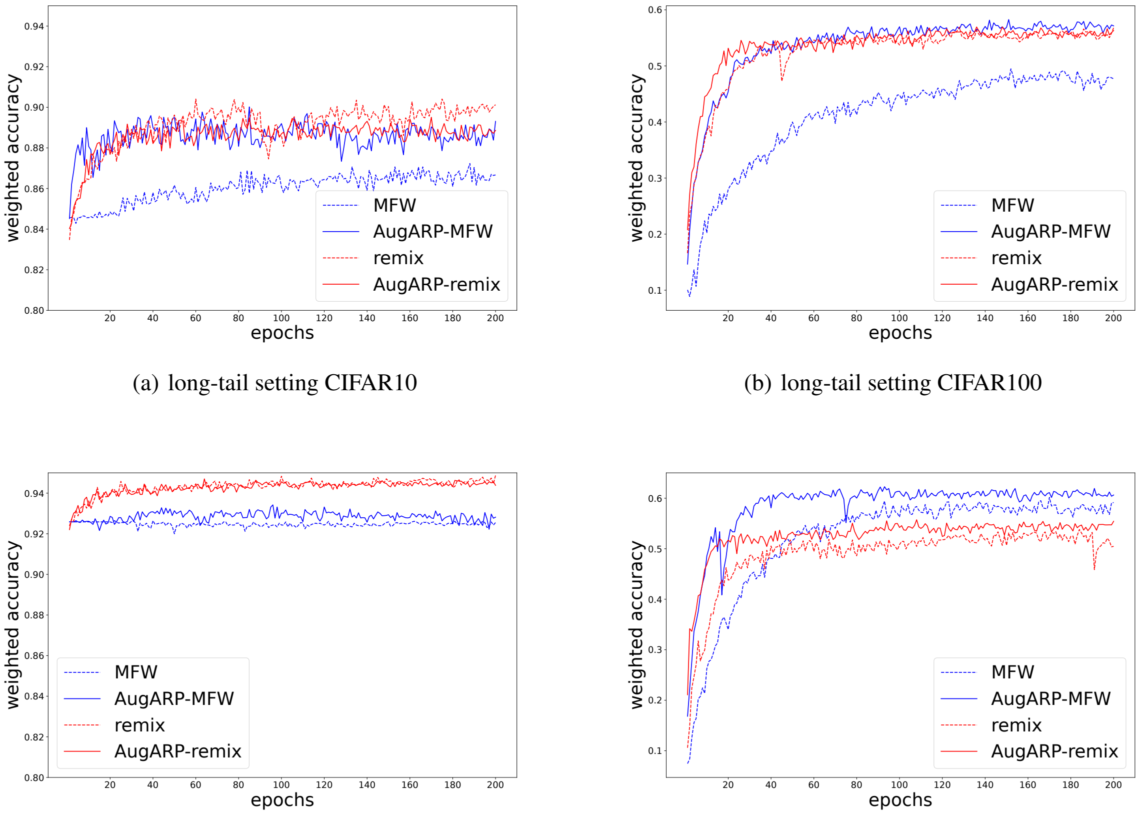
The weighted accuracy is shown in Table 1. Note that traditional long-tail learning methods such as LDAM and focal loss do not perform well in our setting because our proposed setting is different from traditional long-tail learning problems. This result also shows the necessity of AugARP algorithm in the proposed new scenario. On both long-tailed and step-setting CIFAR100 datasets, the AugARP-MFW outperforms other methods. On CIFAR10 dataset, the

[1] https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset

[2] The code is released on https://www.lamda.nju.edu.cn/code_AugARP.ashx

method	CIFAR10		CIFAR100	
	long-tail	step	long-tail	step
ERM	0.6639	0.8020	0.4443	0.5071
LDAM (Cao et al. 2019)	0.8381	0.9258	0.0757	0.0926
Focal (Lin et al. 2020)	0.8325	0.9203	0.2854	0.3416
LDAM-DRW (Cao et al. 2019)	0.8046	0.6852	0.4098	0.4900
MFW (Ye, Zhan, and Chao 2021)	0.8667	0.9255	0.4769	0.5916
Remix (Chou et al. 2020)	**0.9013**	**0.9488**	0.5682	0.5047
AugARP-MFW	0.8930	0.9282	**0.5714**	**0.6064**
AugARP-Remix	0.8886	0.9438	0.5631	0.5544

Table 1: The weighted accuracy of CIFAR10/100 datasets.



(a) long-tail setting CIFAR10

(b) long-tail setting CIFAR100

(c) step setting CIFAR10

(d) step setting CIFAR100

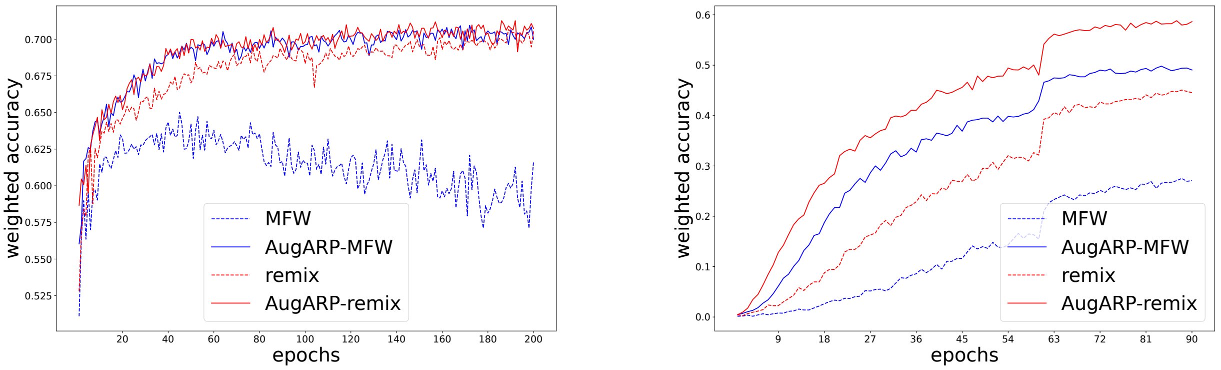
Figure 3: Testing weighted accuracy w.r.t. training epochs on CIFAR10/100 datasets.

weighted accuracy of AugARP method is just slightly behind the weighted accuracy of the best method.

The weighted accuracy of (AugARP) mixup methods is further shown in Fig. 3. The figure shows that AugARP method not only improves the weighted accuracy of Remix and MFW methods but also speeds up the convergence process. This result further validates the effectiveness of AugARP in cost-aware long-tailed learning. The result also indicates that our proposed AugARP method has a better performance on step-setting compared with the long-tailed setting.

Experiment Results on 2023 Amazon Sales Dataset

In this subsection, we conduct the experiment on Amazon Products Sales 2023 dataset, which is an open-source dataset from real-world business. This dataset includes product information over 7 super-classes and 142 classes, and the features include product images, ratings, and prices. In this experiment, we consider the product image classification task, and let the price of each product be the instance-specific misclassification cost. Similar to the experiment settings of CIFAR10/100, we use Resnet-32 for the Amazon Products



(a) Amazon

(b) iNaturalist

Figure 4: Testing weighted accuracy w.r.t. training epochs on Amazon and iNaturalist datasets.

method	Amazon	iNaturalist
ERM	0.6560	0.5077
LDAM	0.2513	0.0011
Focal	0.6342	0.0011
LDAM-DRW	0.5962	0.4620
MFW	0.6168	0.2707
Remix	0.7054	0.4453
AugARP-MFW	0.6999	0.4905
AugARP-Remix	**0.7073**	**0.5865**

Table 2: Comparison on final testing weighted accuracy on Amazon and iNaturalist datasets.

Sales dataset and train 200 epochs for each method. Every algorithm repeats 3 times and the average weighted accuracy is presented. The results of weighted accuracy of both proposed methods and comparison methods are shown in Table 2. According to the result, AugARP-Remix outperforms other methods, which verifies the effectiveness of the proposed AugARP framework on real-world datasets. Fig. 4a shows the weighted accuracy of AugARP methods and comparison mixup methods. The result also verifies that our proposed AugARP method can improve the performance of mixup algorithms in real-world scenarios.

Experiment Results on iNaturalist Dataset

In this subsection, we experiment with the iNaturalist 2018 dataset. The iNaturalist is a large-scale dataset containing over 8,000 classes of 14 super-categories. Similar to the settings of other datasets, the classes from 1 head super-category and 6 tail super-category are set as important classes with misclassification costs of 100 while other classes with misclassification costs of 1. We use Resnet-50 for this experiment and train 90 epochs. As the training process on the large-scale dataset is very time-consuming, we only run every algorithm 1 time to report the weighted accuracy results.

Fig. 4b shows the weighted accuracy of our proposed AugARP Mixup method and comparison Mixup methods. The result proves that our proposed AugARP method can significantly improve the weighted accuracy of Mixup augmentation algorithms. Table 2 refers to further comparison results. Under the settings we consider, other long-tail learning algorithms such as LDAM and focal loss do not work properly, and our proposed AugARP method can achieve good performance. It is also worth noting that, compared with other datasets, our algorithm can significantly outperform other algorithms on the CIFAR100 and iNaturalist datasets, and the characteristics of these two datasets are that they both have a super-class structure. When setting the weight parameters of each class, we set the categories under each super-class to have similar weights. This result is also consistent with the assumptions our proposed AugARP algorithm relies on.

Limitations and Conclusion

Although the existing research assumes the importance of each class is the same in the long-tail learning task, in real-world applications there may exist some tail classes with higher values. In this paper, we propose a novel cost-aware long-tail learning framework AugARP which can adaptively find out the regions of high-value classes implicitly and then perform data augmentation with the region information. We provide both theoretical and empirical analysis to validate the effectiveness of AugARP. As a first step of cost-aware long-tailed learning, there are some limitations in our paper: (1) In the process of `PARTITION`, we use an attention module to divide the region implicitly. However, it may not be the most effective structure. Further research on this issue is necessary to improve the performance of cost-aware long-tailed learning algorithms. (2) Mixup methods used in our proposed algorithm may not be the most effective way to combine the implicit region partition information into the data augmentation process. The best data augmentation method that can take full advantage of region partitions is still waiting to be researched. Despite there are some limitations in our current work, we believe this paper can inspire further research on our proposed cost-aware long-tail learning scenario.

Appendix A: Proof of Theorem 1

The excess risk of $\widehat{h}$ can be decomposed and bounded as follows:

$$\begin{aligned}
&R(\widehat{h};\mathcal{D}) - R(h^*;\mathcal{D}) \\
=&\sum_{k=1}^{K} p_k[R(\widehat{h};\mathcal{D}_k) - R(h^*;\mathcal{D}_k)] \\
\leq&\sum_{k=1}^{K} p_k[R(\widehat{h};\mathcal{D}_k) - R(h_k^*;\mathcal{D}_k)],
\end{aligned}$$

The second equality is due to that h_k^* is optimal for margin risk on the region and the margin error is always larger than the 0-1 classification error. Thus we only need to further bound $R(\widehat{h};\mathcal{D}_k) - R(h_k^*;\mathcal{D}_k)$ on each region. We further have that

$$\begin{aligned}
&R(\widehat{h};\mathcal{D}_k) - R(h_k^*;\mathcal{D}_k) \leq R_\rho(\widehat{h};\mathcal{D}_k) - R_\rho(h_k^*;\mathcal{D}_k) \\
=&R_\rho(\widehat{h};\mathcal{D}_k) - \widehat{R}_\rho(\widehat{h};D_k) + (\widehat{R}_\rho(\widehat{h};D_k) - \widehat{R}_\rho(h_k^*;D_k)) \\
&\quad + \widehat{R}_\rho(h_k^*;D_k) - R_\rho(h_k^*;\mathcal{D}_k) \\
\leq&R_\rho(\widehat{h};\mathcal{D}_k) - \widehat{R}_\rho(\widehat{h};D_k) + \widehat{R}_\rho(h_k^*;D_k) - R_\rho(h_k^*;\mathcal{D}_k) \\
\leq&2(R_\rho(\widehat{h};\mathcal{D}_k) - \widehat{R}_\rho(\widehat{h};D_k)).
\end{aligned}$$

To proceed, we introduce the following lemma, which is the generalization of classical Bernstein's inequality to the bounded difference situation, assuming that the random function to estimate follows from the summation form. The proof is similar to Lemma A.4 in (Ding et al. 2022) by combining with the key observation that under this formulation of f, the sample variance on g is always the upper bound of the variance of the empirical sample mean on g.

Lemma 1. *Assume that there is a real-valued function $f : \mathcal{X}^N \mapsto [0,1]$ such that f has the formulation of $f(X_1, X_2, \ldots, X_N) = \frac{1}{N}\sum_{i=1}^{N} g(X_i)$, in which $g : \mathcal{X} \to [0,1]$ is also a real-valued function bounded in $[0,1]$. Furthermore, denote the variance of g induced by the sample distribution as V, then with probability at least $1-\delta$ over random draw of i.i.d. sample $X_1, X_2, \ldots, X_N$, the following result holds*

$$\begin{aligned}
&f(X_1, X_2, \cdots, X_N) - \mathbb{E}[f(X_1, X_2, \cdots, X_N)] \\
\leq&\sqrt{\frac{2V\log\frac{3}{\delta}}{N}} + \frac{\log\frac{3}{\delta}}{3N}.
\end{aligned}$$

From Lemma 1, we can obtain the Bernstein-type generalization of the classical generalization bound on Rademacher complexity (e.g. Theorem 3.3 of (Mohri, Rostamizadeh, and Talwalkar 2018)), by the substitution of the original McDiarmid's inequality with Lemma 1 in the proof.

Lemma 2. *Let f follow the same formulation in Lemma 1. Denote by $\mathcal{G}$ be the function space of g. Then with probability at least $1-\delta$ over random draw of i.i.d. sample $X_1, X_2, \ldots, X_N$, the following result holds*

$$\begin{aligned}
\mathbb{E}[f(X)] \leq f(X_1, \ldots, X_N) + 2\mathfrak{R}(\mathcal{G}) + \sqrt{\frac{2V\log\frac{3}{\delta}}{N}} \\
+ \frac{\log\frac{3}{\delta}}{3N},
\end{aligned}$$

in which $\mathfrak{R}(\mathcal{G})$ is the Rademacher complexity of function class $\mathcal{G}$.

Set confidence parameter as δ/K we have that for $\forall k \in [K]$, with probability at least $1-\delta/K$,

$$\begin{aligned}
&R_\rho(\widehat{h};\mathcal{D}_k) - \widehat{R}_\rho(\widehat{h};D_k) \\
\leq&2\mathfrak{R}_{N_k}(\mathcal{H}_{\mathrm{CErr}_\rho}) + \sqrt{\frac{2V\log\frac{3K}{\delta}}{N_k}} + \frac{\log\frac{3K}{\delta}}{3N_k},
\end{aligned}$$

in which $\mathcal{H}_{\mathrm{CErr}_\rho}$ is the function class induced by $\mathcal{H}$ and CErr_ρ. Then following the similar proof steps to Theorem 1 of (Kuznetsov, Mohri, and Syed 2014), we have

$$\mathfrak{R}_{N_k}(\mathcal{H}_{\mathrm{CErr}_\rho}) \leq \frac{4LL_\sigma}{\rho}\mathfrak{R}_{N_k}(\mathcal{W}),$$

in which $\mathcal{W} = \{w : w \in W\}$ is the function set of all linear prediction heads. The final task is to bound $\mathfrak{R}_{N_k}(\mathcal{H}_{\mathrm{CErr}_\rho})$. Denote by ν the Rademacher random variables:

$$\begin{aligned}
&\mathfrak{R}_{N_k}(\mathcal{H}_{\mathrm{CErr}_\rho}) \\
=&\frac{1}{N_k}\mathbb{E}_{\mathcal{D}_{\phi_k},\nu}\big[\sup_{\|w\|\leq\Lambda}\langle w, \sum_{i=1}^{N_k}\nu_i\widehat{\phi}(x_i)\rangle\big] \\
\leq&\frac{1}{N_k}\mathbb{E}_{\mathcal{D}_{\phi_k},\nu}\big[\sup_{\|w\|\leq2\Lambda}\langle w, \sum_{i=1}^{N_k}\nu_i(\widehat{\phi}(x_i) - \mathbb{E}_{\mathcal{D}_{\phi_k}}[\widehat{\phi}(x)])\rangle\big] \\
\leq&\frac{2\Lambda}{N_k}\mathbb{E}_{\mathcal{D}_{\phi_k},\nu}\Big[\sup_{\|w\|\leq2\Lambda}\|w\|\big[\|\sum_{i=1}^{N_k}\nu_i(\widehat{\phi}(x_i) - \mathbb{E}_{\mathcal{D}_{\phi_k}}[\widehat{\phi}(x)])\|\big]\Big] \\
\leq&\frac{2\Lambda}{N_k}\mathbb{E}_{\mathcal{D}_{\phi_k},\nu}\big[\|\sum_{i=1}^{N_k}\nu_i(\widehat{\phi}(x_i) - \mathbb{E}_{\mathcal{D}_{\phi_k}}[\widehat{\phi}(x)])\|\big] \\
\leq&\frac{2\Lambda}{N_k}\Big[\mathbb{E}_{\mathcal{D}_{\phi_k},\nu}\big[\|\sum_{i=1}^{N_k}\nu_i(\widehat{\phi}(x_i) - \mathbb{E}_{\mathcal{D}_{\phi_k}}[\widehat{\phi}(x)])\|\big]^2\Big]^{\frac{1}{2}} \\
=&\frac{2\Lambda}{N_k}\Big[\mathbb{E}_{\mathcal{D}_{\phi_k}}\big[\sum_{i=1}^{N_k}\|(\widehat{\phi}(x_i) - \mathbb{E}_{\mathcal{D}_{\phi_k}}[\widehat{\phi}(x)])\|^2\big]\Big]^{\frac{1}{2}} \\
=&\frac{2\Lambda}{\sqrt{N_k}}\Big[\mathbb{E}_{\mathcal{D}_{\phi_k}}\big[\|(\widehat{\phi}(x) - \mathbb{E}_{\mathcal{D}_{\phi_k}}[\widehat{\phi}(x)])\|^2\big]\Big],
\end{aligned}$$

in which the first inequality is due the compactness of $\mathcal{W}$, the second inequality is due to Cauchy-Schwarz inequality, the fourth inequality is due to Jensen's inequality, the second equality is due to that ν are zero-mean i.i.d. random variables. Then we arrive at the final result.

Appendix B: Analysis of the Correlation Assumption

The correlation assumption between sample features and misclassification costs plays a very important role in our

	Spearman's coefficient of correlation	P value
Euclidean distance without embedding	0.0348	$< 10^{-5}$
Cosine distance without embedding	0.0437	$< 10^{-5}$
Euclidean distance with embedding	0.3323	$< 10^{-5}$
Cosine distance with embedding	0.4786	$< 10^{-5}$

Table 3: The Spearman's coefficient of correlation between the feature similarity and cost similarity

proposed approach. Thus it is necessary to examine whether the assumption is consistent with reality. In this section, we conduct a brief analysis of the assumption on the Amazon Product Sales dataset.

On the Amazon Product Sales dataset, the feature and misclassification cost of each good item is defined as the pictures of the good and the price of the good. We calculate Spearman's rank coefficient of correlation (Fieller, Hartley, and Pearson 1957) between the misclassification cost gap and feature similarity of sample pairs on the dataset. For a sample pair $(\mathbf{x}_i, c_i), (\mathbf{x}_j, c_j)$, the misclassification cost gap is defined as $(c_i - c_j)^2$, and the feature similarity is defined by two ways: the Euclidean distance $||\mathbf{x}_i - \mathbf{x}_j||$, and the cosine distance $1 - \mathbf{x}_i^T \mathbf{x}_j / (||\mathbf{x}_i|| \cdot ||\mathbf{x}_j||)$. Considering the influence of the feature embedding process, we calculate the result for both original features and embedded features. The result is shown as Table 1. We use the Resnet-32 trained with 100 epochs for feature embeddings.

Note that all P values are very small, which indicates that all correlation results are significant. For the original features without feature embedding, the value of the correlation coefficient is relatively small, indicating that the correlation in this case is weak. However, for the features after feature embedding, the value of the correlation coefficient is relatively large, which indicates that there is indeed a certain correlation between feature similarity and cost similarity after feature embedding. This result confirms that our hypothesis is reasonable and consistent with real-world scenarios.

Appendix C: Attention Module

In our proposed AugARP approach, we introduce an individual attention module after the feature-extracting network to divide the embedding space implicitly. Here we use the most basic attention module:

$$\mathtt{Att} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

in which Q represents the embedded batch input, K and V represents the keys and values parameter matrices.

In our work, the attention mechanism is used to measure the similarity between samples, including the similarity between sample features and the similarity between sample importance. In order to achieve this goal, we introduce the following regular terms to the loss function during the training process to train the attention module:

$$\frac{\lambda_1}{m^2}\sum_{i,j}(\mathtt{Att}_{i,j} - s_{i,j})^2 + \frac{\lambda_2}{m^2}\sum_{i,j}(\mathtt{Att}_{i,j} - c_{i,j})^2$$

in which $s_{i,j}$ represents the cosine similarity of embedded samples, and $c_{i,j}$ represents the similarity of (normalized) misclassification costs, m is batch size.

Broader Impact

In cost-aware long-tailed classification, focusing learning on those high-cost minority classes is essential. Similar to previous long-tailed classification studies, as the importance of minorities is emphasized in our setting, we believe that the AugARP method proposed in this paper can help to improve social good and fairness. Since the cost-aware scenario appears in many real-world applications, AugARP has the potential significance to mitigate the safety risk in them. For example, among different kinds of rare social behaviors, AugARP may help to distinguish truly malicious ones that could lead to huge hazards, so that public resources can be efficiently used.

On the other hand, under the cost-aware setting, incorrect cost setting and estimation might also lead to negative effects on minorities, who are truly important but are assigned low costs mistakenly. We believe that ensuring the proper assignment and estimation of costs under cost-aware long-tailed classification is an essential problem for the next step of research.

Acknowledgements

This work is supported by National Science Foundation of China (62206245, 62376118). We also thank the anonymous reviewers for their constructive suggestions.

References

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106(C): 249–259.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 1567–1578.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1): 321–357.

Chou, H.-P.; Chang, S.-C.; Pan, J.-Y.; Wei, W.; and Juan, D.-C. 2020. Remix: Rebalanced Mixup. In *Computer Vision – ECCV 2020 Workshops*, 95–110.

Cortes, C.; DeSalvo, G.; Gentile, C.; Mohri, M.; and Zhang, N. 2019. Region-Based Active Learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2801–2809.

Cortes, C.; DeSalvo, G.; Gentile, C.; Mohri, M.; and Zhang, N. 2020. Adaptive Region-Based Active Learning. In *Proceedings of the 37th International Conference on Machine Learning*, 2144–2153.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9260–9269.

Ding, Y.-X.; Wu, X.-Z.; Zhou, K.; and Zhou, Z.-H. 2022. Pre-Trained Model Reusability Evaluation for Small-Data Transfer Learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 37389–37400.

Elkan, C. 2001. The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, 973–978.

Fieller, E. C.; Hartley, H. O.; and Pearson, E. S. 1957. Tests for Rank Correlation Coefficients. I. *Biometrika*, 44(3-4): 470–481.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *Proceedings of the 8th International Conference on Learning Representations*.

Kim, B.; and Kim, J. 2020. Adjusting decision boundary for class imbalanced learning. *IEEE Access*, 8: 81674–81685.

Kuznetsov, V.; Mohri, M.; and Syed, U. 2014. Multi-class deep boosting. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 2501–2509.

Li, Y.; Wang, T.; Kang, B.; Tang, S.; Wang, C.; Li, J.; and Feng, J. 2020. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10991–11000.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327.

Liu, X.-Y.; Wu, J.; and Zhou, Z.-H. 2009. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2): 539–550.

Liu, X.-Y.; and Zhou, Z.-H. 2006. The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study. In *Proceedings of the 6th International Conference on Data Mining*, 970–974.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2537–2546.

Mazurowski, M. A.; Habas, P. A.; Zurada, J. M.; Lo, J. Y.; Baker, J. A.; and Tourassi, G. D. 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2-3): 427–436.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. *Foundations of machine learning*. MIT press.

Oksuz, K.; Cam, B. C.; Kalkan, S.; and Akbas, E. 2020. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10): 3388–3415.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11662–11671.

Wei, T.; and Li, Y.-F. 2018. Does Tail Label Help for Large-Scale Multi-Label Learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2847–2853.

Ye, H.-J.; Chen, H.-Y.; Zhan, D.-C.; and Chao, W.-L. 2020. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*.

Ye, H.-J.; Zhan, D.-C.; and Chao, W.-L. 2021. Procrustean Training for Imbalanced Deep Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 92–102.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, S.; Li, Z.; Yan, S.; He, X.; and Sun, J. 2021. Distribution Alignment: A Unified Framework for Long-Tail Visual Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2361–2370.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9719–9728.

Zhou, Z.-H. 2021. *Machine learning*. Springer Nature.

Zhou, Z.-H.; and Liu, X.-Y. 2010. On Multi-Class Cost-Sensitive Learning. *Computational Intelligence*, 26(3): 232–257.