Dynamic Curriculum Learning for Imbalanced Data Classification

Yiru Wang *, Weihao Gan *, Jie Yang, Wei Wu, Junjie Yan
SenseTime Group Limited
{wangyiru,ganweihao,yangjie,wuwei,yanjunjie}@sensetime.com

Abstract

Human attribute analysis is a challenging task in the field of computer vision. One of the significant difficulties is brought from largely imbalance-distributed data. Conventional techniques such as re-sampling and cost-sensitive learning require prior-knowledge to train the system. To address this problem, we propose a unified framework called Dynamic Curriculum Learning (DCL) to adaptively adjust the sampling strategy and loss weight in each batch, which results in better ability of generalization and discrimination. Inspired by curriculum learning, DCL consists of two-level curriculum schedulers: (1) sampling scheduler which manages the data distribution not only from imbalance to balance but also from easy to hard; (2) loss scheduler which controls the learning importance between classification and metric learning loss. With these two schedulers, we achieve state-of-the-art performance on the widely used face attribute dataset CelebA and pedestrian attribute dataset RAP.

1. Introduction

Human attribute analysis, including facial characteristics and clothing categories, has facilitated the society in various aspects, such as tracking and identification. However, different from the general image classification problem like ImageNet challenge [28], human attribute analysis naturally involves largely imbalanced data distribution. For example, when collecting the face data of attribute 'Bald', most of them would be labeled as 'No Bald' and its imbalanced ratio to the 'Bald' class would be relatively high. Training the classification model with equal importance for samples in different classes may result in a bias to the majority class of the data and poor accuracy for the minority class. Therefore, it is of great importance to handle the imbalanced data learning problem, especially in human attribute analysis.

Impressive results have been achieved for the general imbalanced data learning in the past years. One intuitive motivation is resampling [2, 9, 13, 16, 36, 15], which either oversamples the minority class data or downsamples the majority class data, to balance the data distribution. However, oversampling could easily cause overfitting problem due to repeatedly visiting duplicated minority samples, while downsampling may discards much useful information in the majority samples. Another kind of approach called cost-sensitive learning is also exploited to handle the imbalanced data learning problem, which directly imposes heavier cost on the misclassified minority class [48, 50, 55, 57] . However, it is difficult to determine the exact cost for different samples in various distributions. Hand et al. [14] proposed a batch-wise method that selects part of the majority samples and increases the weight of minority samples to match a pre-defined target distribution. Besides the standard cross entropy classification loss, Dong et al. [7, 8] proposed to add another class rectification loss (CRL) to avoid the dominant effect of majority classes. A specific metric is proposed for imbalanced datasets by above methods. For the general classification problem, **class-biased accuracy** is defined as the number of correctly predicted samples divided by the number of the whole test data. While for imbalanced data classification, **class-balanced accuracy** is defined as the average of the accuracy in each class for evaluation.

Our proposed Dynamic Curriculum Learning (DCL) method is motivated by the following two considerations. (1) Sampling is an acceptable strategy for the problem, but keeping targeting at a balanced distribution in the whole process would hurt the generalization ability, particularly for a largely imbalanced task. For example, in the early stage of learning with balanced target distribution, the system discards lots of majority samples and emphasizes too much on the minority samples, tending to learn the valid representation of the minority class but the bad/unstable representation of the majority one. However, what we expect is to make the system first learn the appropriate general representations for both of the classes on the target attributes and then classify the samples into correct labels, which results in a favorable balance between the class bias accuracy and class balanced accuracy. (2) It is reasonable to combine cross entropy loss (CE) and metric learning loss

*Equal contribution.

(ML) since the appropriate feature representation could be helpful for classification. However, we think those two components contribute to different emphasis. Treating them equally in the training process cannot fully utilize the discriminative power of deep CNN. Specifically, CE pays more attention to the classification task by assigning specific labels, while ML focuses more on learning a soft feature embedding to separate different samples in feature space without assigning labels. Similarly to the previous point, we expect the system first to learn the appropriate feature representation and then classify the samples into the correct labels.

In the spirit of the curriculum learning [1], we propose Dynamic Curriculum Learning (DCL) framework for imbalanced data learning. Specifically, we design two-level curriculum schedulers: (1) sampling scheduler: it aims to find the most meaningful samples in one batch to train the model dynamically from imbalanced to balanced and from easy to hard; (2) loss scheduler: it controls the learning weights between classification loss and metric learning loss. These two components can be defined by the scheduler function, which reflects the model learning status. To summarize our contributions:

- For the first time, we introduce the curriculum learning idea into imbalanced data learning problem. Based on the designed scheduler function, two curriculum schedulers are proposed for dynamic sampling operation and loss backward propagation.
- The proposed DCL framework is a unified representation, which can generalize to several existing state-of-the-art methods with corresponding setups.
- We achieve the new state-of-the-art performance on the commonly used face attribute dataset CelebA [35] and pedestrian attribute dataset RAP [31].

2. Related Work

Imbalanced data learning. There are several groups of methods trying to address the imbalanced learning problem in literature. (1) Data-level: considering the imbalanced distribution of the data, one intuitive way to do is resampling the data [2, 9, 13, 16, 36, 38, 17, 10] into a balanced distribution, which could oversample the minority class data and downsample the majority class data. One advanced sampling method called SMOTE [2, 3] augments artificial examples created by interpolating neighboring data points. Some extensions of this technique were proposed [13, 36]. However, oversampling can easily cause overfitting problem due to repeatedly visiting duplicated minority samples. While downsampling usually discards many useful information in majority samples. (2) Algorithm-level: cost-sensitive learning aims to avoid above issues by directly imposing a heavier cost on misclassifying the minority class [48, 50, 55, 57, 53, 49]. However, how to determine the cost representation in different problem settings or environments is still an open question. Besides of the cost-sensitive learning, another option is to change the decision threshold during testing, which is called threshold-adjustment technique [5, 54, 57]. (3) Hybrid: this is an approach that combines multiple techniques from one or both abovementioned categories. Widely used example is ensembling idea. EasyEnsemble and BalanceCascade are methods that train a committee of classifiers on undersampled subsets [34]. SMOTEBoost, on the other hand, is a combination of boosting and SMOTE oversampling [4]. Some methods like [37, 26, 44, 51, 56, 39] also pays attention to the noisy samples in the imbalanced dataset.

Deep imbalanced learning. Recently, several deep methods have been proposed for imbalanced data learning [7, 8, 14, 19, 20, 22, 24, 25, 45, 57, 42, 6]. One major direction is to integrate the sampling idea and cost-learning into an efficient end-to-end deep learning framework. Jeatrakul et al. [22] treated the Complementary Neural Network as an under-sampling technique, and combined it with SMOTE-based over-sampling to rebalance the data. Zhou et al. [57] studied data resampling for training cost-sensitive neural networks. In [24, 6], the cost-sensitive deep features and the cost parameter are jointly optimized. Oquab et al. [38] resampled the number of foreground and background image patches for learning a convolutional neural network (CNN) for object classification. Hand et al. [14] proposed a selective learning(SL) method to manage the sample distribution in one batch to a target distribution and assign larger weight for minority classes for backward propagation. Another recent direction of the problem involves the metric learning into the system. Dong et al. [7, 8] proposed a class rectification loss (CRL) regularising algorithm to avoid the dominant effect of majority classes by discovering sparsely sampled boundaries of minority classes. More recently, LMLE/CLMLE [19, 20] are proposed to preserve the local class structures by enforcing large margins between intra-class and inter-class clusters.

Curriculum learning. The idea of curriculum learning was originally proposed in [1], it demonstrates that the strategy of learning from easy to hard significantly improves the generalization of the deep model. Up to now, works been done via curriculum learning mainly focus on visual category discovery [29, 41], object tracking [47], semi-/weakly-supervised learning [11, 12, 23, 40], etc. [40] proposed an approach that processes multiple tasks in a sequence with sharing between subsequent tasks instead of solving all tasks jointly by finding the best order of tasks to be learned. Very few works approach the imbalanced learning. Guo et al. [12] developed a principled learning strategy by leveraging curriculum learning in a weakly supervised framework, with the goal of effectively learning from imbalanced data.

3. Method

We propose a Dynamic Curriculum Learning (DCL) framework for imbalanced data classification problem, consisting of two-level **curriculum schedulers**. The first one is a sampling scheduler of which the key idea is to find the most significant samples in one batch to train the model dynamically making data distribution from imbalanced to balanced and from easy to hard. This scheduler determines the sampling strategy for the proposed Dynamic Selective Learning (DSL) loss function. The second one is the loss scheduler, which controls the learning importance between two losses: the DSL loss and the metric learning loss (triplet loss). Therefore, in the early stage of the training process, the system focuses more on the soft feature space embedding, while later on, it pays more attention to the task of classification.

3.1. Scheduler Function Design

Most of the traditional curriculum learning methods manually define different training strategies. While in our proposed DCL framework for imbalanced data learning, we formulate the key idea of curriculum scheduling with different groups of functions, as we called **Scheduler Function**. We show the semantic interpretation for those functions.

The scheduler function $SF(l)$ is a function which returns value monotonically decreasing from 1 to 0 with the input variable l, which represents the current training epoch. It reflects the model learning status and measures the curriculum learning speed. We explore several function classes as following (illustrated in Figure 1):

- Convex function: indicating the learning speed from slow to fast. For example:

$$SF_{cos}(l) = cos(\frac{l}{L} * \frac{\pi}{2}) \tag{1}$$

- Linear function: indicating the constant learning speed. For example:

$$SF_{linear}(l) = 1 - \frac{l}{L} \tag{2}$$

- Concave function: indicating the learning speed from fast to slow. For example:

$$SF_{exp}(l) = \lambda^l \tag{3}$$

- Composite function: indicating the learning speed from slow to fast and then slow again. For example:

$$SF_{composite}(l) = \frac{1}{2}cos(\frac{l}{L}\pi) + \frac{1}{2} \tag{4}$$

where L refers to expected total training epochs and λ is an independent hyperparameter that in the range of $(0, 1)$.

Different classes of $SF(l)$ represent different curriculum learning styles. Based on the above-introduced scheduler functions, we propose Dynamic Curriculum Learning framework for imbalanced data classification.

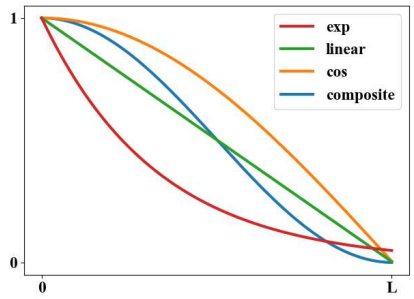

Figure 1. Four types of designed scheduler functions.

3.2. Sampling Scheduler

Sampling is one of the most commonly used techniques to deal with imbalanced data learning. In this section, we introduce the proposed Dynamic Selective Learning (DSL) component, which is based on our sampling scheduler. The sampling scheduler dynamically adapts the target distribution in a batch from imbalanced to balanced during the training process.

Explicitly, for each attribute, we define j^{th} element of the data distribution D as the number of j^{th} class samples divided by the number of minority samples (the least one). Sorting them in ascending order, then we have:

$$D = 1 : \frac{\#C_1}{\#C_{min}} : \frac{\#C_2}{\#C_{min}} : ... : \frac{\#C_{K-1}}{\#C_{min}} \tag{5}$$

where K is the number of classes and $\#C_i$ is the number of samples in class i. Each attribute has its training distribution D_{train}, which is a global statistic.

Sampling scheduler determines the target data distribution of the attributes in each batch. Initially, the target distribution of one attribute $D_{target}(0)$ in a batch is set to D_{train}, which is imbalanced distributed. During the training process, it gradually transfers to a balanced distribution with the following function (each element is powered by $g(l)$):

$$D_{target}(l) = {D_{train}}^{g(l)} \tag{6}$$

where l refers to current training epoch and $g(l)$ is the sampling scheduler function, which can be any choice in Section 3.1. According to target distribution $D_{target}(l)$, the majority class samples are dynamically selected and the minority class samples are re-weighted in different epochs to confirm different target distributions in one batch. Therefore, the DSL loss is defined as:

$$\mathcal{L}_{\text{DSL}} = -\frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{N_j} w_j * \log\left(p(y_{i,j} = \bar{y}_{i,j}|\mathbf{x}_{i,j})\right) \tag{7}$$

$$w_j = \begin{cases} \frac{D_{target,j}(l)}{D_{current,j}} & \text{if } \frac{D_{target,j}(l)}{D_{current,j}} \geq 1 \\ 0/1 & \text{if } \frac{D_{target,j}(l)}{D_{current,j}} < 1 \end{cases} \tag{8}$$

where N is batch size, N_j is the number of samples of j^{th} class in current batch, M is number of classes, $\bar{y}_{i,j}$ is the ground truth label. w_j is the cost weight for class j. $D_{target,j}(l)$ is the j^{th} class target distribution in current epoch l. $D_{current,j}$ is the j^{th} class distribution in current batch before sampling. If $\frac{D_{target,j}(l)}{D_{current,j}} < 1$, we sample $\frac{D_{target,j}(l)}{D_{current,j}}$ percentage of j^{th} class data with original weight 1 and the remainings with 0. If not, then j^{th} class is a minority class and a larger weight is assigned to the samples.

With different sampling scheduler functions (four types in the previous section), the batch target distribution changes from the training set biased distribution to balanced distribution. At the beginning epoch, $g(0) = 1$, the target distribution D equals to the train set distribution; in other words, the real-world distribution. At the final epoch, $g(l)$ is close to 0, so all the element in target distribution D is close to 1 (power of 0). In other words, it is a balanced distribution.

The learning rate is usually set conforming to a decay function. At the early stage of the training process, with a large learning rate and biased distribution, the curriculum scheduler manages the model to learn more on whole training data. Usually, the system learns lots of easy samples in this stage. Going further with the training process, the target distribution is gradually getting balanced. With the selected majority samples and re-weighted minority samples, the system focuses more on the harder situation.

3.3. Metric Learning with Easy Anchors

Besides of the loss function$\mathcal{L}_{\text{DSL}}$, we also involve a metric learning loss to learn a better feature embedding for imbalance data classification.

A typical selection of the metric learning loss is triplet loss, which was introduced by CRL[8] with hard mining. Define the samples with high prediction score on the wrong class as hard samples. Then we build triplet pairs from the anchors and some hard positive and negative samples. The loss function in CRL is defined as following:

$$\mathcal{L}_{\text{crl}} = \frac{\sum_T \max\left(0,\ m_j + d(\mathbf{x}_{all,j}, \mathbf{x}_{+,j}) - d(\mathbf{x}_{all,j}, \mathbf{x}_{-,j})\right)}{|T|} \tag{9}$$

where m_j refers to the margin of class j in triplet loss and $d(\cdot)$ denotes the feature distance between two samples. In current batch, $\mathbf{x}_{all,j}$ represents all the samples in class j, $\mathbf{x}_{+,j}$ and $\mathbf{x}_{-,j}$ represents positive samples and negative samples respectively. T refers to the number of triplet pairs. In CRL[8], all the minority class samples are selected as anchors.

We define **easy sample** as the correctly predicted sample. Choosing all the minority samples as anchors is not stable for model to learn, since it may cause problems such as pulling easy positive samples to the negative side. Examples are illustrated in Figure 2.

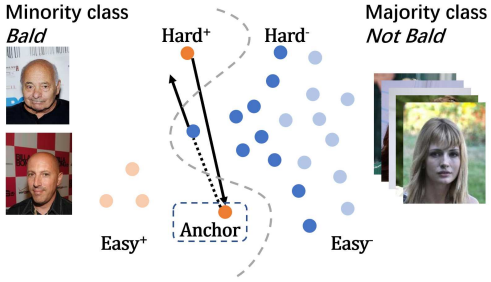

Figure 2. This figure visualizes a case of Triplet Loss in CRL[8] that hard positive sample is chosen as the anchor. Assuming minority class as the positive class, the triplet pair shown in the figure is trying to push both the positive sample and the negative sample across the border, which is pushing the positive sample closer to the negative side. It can cause the features of positive samples to be more chaotic.

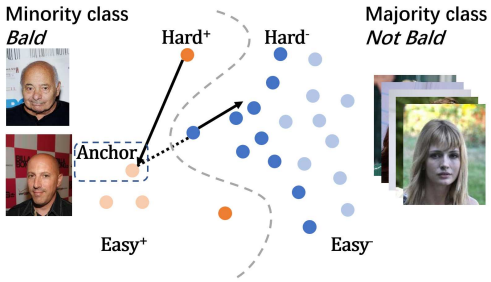

Figure 3. This figure visualizes a case of our proposed Triplet Loss with only easy positive samples as the anchor. Since easy positive samples' features can be grouped easily, the hard positive sample can be pulled closer to all the easy positive samples. Our proposed method can avoid the situation in Figure 2.

We propose a method to improve the sampling operation of Triplet loss with Easy Anchors $\mathcal{L}_{\text{TEA}}$, defined as follow:

$$\mathcal{L}_{\text{TEA}} = \frac{\sum_T \max\left(0,\ m_j + d(\mathbf{x}_{easy,j}, \mathbf{x}_{+,j}) - d(\mathbf{x}_{easy,j}, \mathbf{x}_{-,j})\right)}{|T|} \tag{10}$$

where $\mathbf{x}_{easy,j}$ refers to easy minority samples in class j, others are similar to equation 9. Easy anchors are defined as high-confident correctly predicted minority samples. The number of hard positives, hard negatives and easy anchors to be selected is determined by the hyper-parameter k.

With $\mathcal{L}_{\text{TEA}}$ loss, only easy samples in minority class are chosen as anchors, which pulls the hard positive samples closer and pushes hard negative samples further. As illustrated in Figure 3. Different from CRL choosing all minority samples as anchors to make rectification on feature space, our proposed method selects easy anchors based on the result of the classifier and pull all the samples to well-

classified side. Also, we adopt the hard sample mining for those selected easy anchors to build the triplet loss.

3.4. Loss Scheduler

To train the model better, we analyze the different characteristics of the two proposed losses. Generally speaking, triplet loss targets at learning a soft feature embedding to separate different samples in feature space without assigning labels, while cross entropy loss aims to classify the samples by assigning specific labels.

Particularly for imbalanced data learning, what we want is that the system first learns an appropriate feature representation then benefits the classification. Therefore, in order to fully utilize these two properties, we design a loss curriculum scheduler $f(l)$ to manage these two losses.

Even though we can choose any one of the schedule functions in Section 3.1, we use the composite function (Equation 4) as an example here. The model learns with the following scheduler:

$$\mathcal{L}_{\text{DCL}} = \mathcal{L}_{\text{DSL}} + f(l) * \mathcal{L}_{\text{TEA}} \tag{11}$$

$$f(l) = \begin{cases} \frac{1}{2}cos(\frac{l}{L}\pi) + \frac{1}{2} + \epsilon & \text{if } l < pL \\ \epsilon & \text{if } l \geq pL \end{cases} \tag{12}$$

where l refers to current training epoch, L refers to expected total training epochs. Small modifications including a hyperparameter p ranging in $[0, 1]$, which is defined as advanced self-learning point. Moreover, ϵ is the self-learning ratio. The reason why we have a non-zero ϵ here is that even though in self-learning stage, the model still needs to maintain the feature structure learned from in the previous stages.

In the early stage of training, a large weight is initialized to the triplet loss $\mathcal{L}_{\text{TEA}}$ for learning soft feature embedding and decreases through time in respect to the scheduler function. In the later stage, the scheduler assigns a small impact on $\mathcal{L}_{\text{TEA}}$ and system emphasizes more on the Dynamic Selective Loss $\mathcal{L}_{\text{DSL}}$ to learn the classification. Finally, when it reaches the self-learning point, no 'teacher' curriculum scheduler is needed. The model automatically finetunes the parameters until convergence.

3.5. Generalization of DCL Framework

To handle the imbalanced data learning problem, we propose the Dynamic Curriculum Learning framework. Revisiting the overall system, DCL consists of two-level curriculum schedulers. One is for sampling $g(l)$, and another is for loss learning $f(l)$. We can find that several state-of-the-art imbalanced learning methods can be generalized from the framework with different setups for the schedulers. The correspondings are listed in Table 1. Selective Learning [14] does not contain metric learning and only uses a fixed target distribution. CRL-I[7] does not contain a re-weight or re-sample operation and only uses a fixed weight for metric learning.

Method	g(x)	f(x)
Cross Entropy	1	0
Selective Learning[14]	0/1	0
CRL-I[7]	1	ϵ
DCL(Ours)	Sampling scheduler	Loss scheduler

Table 1. Generalization of proposed Dynamic Curriculum Learning method to other non-clustering imbalanced learning methods with corresponding setups.

4. Experiments

4.1. Datasets

CelebA [35] is a human facial attribute dataset with annotations of 40 binary classifications. CelebA is an imbalanced dataset, specifically on some attributes, where the sample imbalance level (majority class rate-50%) could be up to 48. The dataset contains 202,599 images from 10,177 different people.

RAP [31] is a richly annotated dataset for pedestrian attribute recognition in real surveillance scenario. It contains 41,585 images from 26 indoor cameras, with 72 different attributes. RAP is a highly imbalanced dataset with the imbalance ratio (minority samples to majority samples) up to 1:1800.

CIFAR-100 [27] is a natural image classification dataset with 32×32 pixels. It contains 50,000 images for training and 10,000 images for testing. It is a balanced dataset with 100 classes. Each class holds the same number of images.

4.2. Evaluation Metric

For CelebA dataset and RAP dataset, following the standard profile, we apply the class-balanced accuracy (binary classification) on every single task, and then compute the mean accuracy of all tasks as the overall metric. It can be formulated as following:

$$mA_i = \frac{1}{2}(\frac{TP_i}{P_i} + \frac{TN_i}{N_i}) \tag{13}$$

$$mA = \frac{\Sigma_{i=1}^{|C|} mA_i}{|C|} \tag{14}$$

where mA_i indicates the class-balanced mean accuracy of the i-th task, with TP_i and P_i indicating the count of predicted true positive samples and positive samples in the ground truth for the i-th task while TN_i and N_i refers to the opposite. $|C|$ is the number of tasks.

For CIFAR-100 dataset, since each class holds the same number of instances, class-balanced accuracy equals to class-biased accuracy.

Table 2. Class-balanced Mean Accuracy (mA) for each class (%) and class imbalance level (majority class rate-50%) of each of the attributes on CelebA dataset. The 1st/2nd best results are highlighted in red/blue.

	Attractive	Mouth Open	Smiling	Wear Lipstick	High Cheekbones	Male	Heavy Makeup	Wavy Hair	Oval Face	Pointy Nose	Arched Eyebrows	Black Hair	Big Lips	Big Nose	Young	Straight Hair	Brown Hair	Bags Under Eyes	Wear Earrings	No Beard	Bangs
Imbalance level	1	2	2	3	5	8	11	18	22	22	23	26	26	27	28	29	30	30	31	33	35
DeepID2(CE) [46]	78	89	89	92	84	94	88	73	63	66	77	83	62	73	76	65	79	74	75	88	91
Over-Sampling [9]	77	89	90	92	84	95	87	70	63	67	79	84	61	73	75	66	82	73	76	88	90
Down-Sampling [9]	78	87	90	91	80	90	89	70	58	63	70	80	61	76	80	61	76	71	70	88	88
Cost-Sensitive [16]	78	89	90	91	85	93	89	75	64	65	78	85	61	74	75	67	84	74	76	88	90
Selective-Learning [14]	81	91	92	93	86	97	90	78	66	70	79	87	66	77	83	72	84	79	80	93	94
CRL-I [7]	83	95	93	94	89	96	84	79	66	**73**	80	**90**	**68**	**80**	84	73	86	80	83	94	95
LMLE [19]	**88**	**96**	**99**	**99**	**92**	**99**	**98**	**83**	68	72	79	92	60	**80**	**87**	73	**87**	73	83	**96**	**98**
CLMLE [20]	**90**	**97**	**99**	**98**	**94**	**99**	**98**	**87**	**72**	**78**	**86**	**95**	66	**85**	**90**	**80**	**89**	**82**	**86**	**98**	**99**
DCL (ours)	83	93	93	95	88	98	92	81	**70**	**73**	**82**	89	**69**	**80**	86	**76**	86	**82**	**85**	95	96

	Blond Hair	Bushy Eyebrows	Wear Necklace	Narrow Eyes	5 o'clock Shadow	Receding Hairline	Wear Necktie	Eyeglasses	Rosy Cheeks	Goatee	Chubby	Sideburns	Blurry	Wear Hat	Double Chin	Pale Skin	Gray Hair	Mustache	Bald		**Average**
Imbalance level	35	36	38	38	39	42	43	44	44	44	44	44	45	45	45	46	46	46	48		
DeepID2(CE) [46]	90	78	70	64	85	81	83	92	86	90	81	89	74	90	83	81	90	88	93		**81.17**
Over-Sampling [9]	90	80	71	65	85	82	79	91	90	89	83	90	76	89	84	82	90	90	92		**81.48**
Down-Sampling [9]	85	75	66	61	82	79	80	85	82	85	78	80	68	90	80	78	88	60	79		**77.45**
Cost-Sensitive [16]	89	79	71	65	84	81	82	91	**92**	86	82	90	76	90	84	80	90	88	93		**81.60**
Selective-Learning [14]	93	85	73	**74**	89	**87**	92	97	90	94	**87**	**94**	**86**	96	**89**	92	94	92	95		**85.93**
CRL-I [7]	95	84	**74**	72	90	**87**	88	96	88	96	**87**	92	85	98	**89**	92	95	**94**	**97**		**86.60**
LMLE [19]	**99**	82	59	59	82	76	90	98	78	95	79	88	59	**99**	74	80	91	73	90		**83.83**
CLMLE [20]	**99**	**88**	69	71	**91**	82	**96**	**99**	86	**98**	85	**94**	72	**99**	87	**94**	**96**	82	95		**88.78**
DCL (ours)	95	**87**	**76**	**79**	**93**	**90**	**95**	**99**	**92**	**97**	**93**	**97**	**93**	**99**	**94**	**96**	**99**	**97**	**99**		**89.05**

4.3. Experiments on CelebA Face Dataset

4.3.1 Implementation Details

Network Architecture We use DeepID2[46] as the backbone for experiments on CelebA for a fair comparison. DeepID2[46] is a CNN of 4 convolution layers. All the experiments listed on table 2 set DeepID2[46] as backbone. The baseline is trained with a simple Cross-Entropy loss. Since CelebA is a multi-task dataset, we set an independent 64D feature layer and a final output layer for each task branch. For each branch, it considers its own current and target distribution and generates single attribute loss (Equation 12). Then we sum them up for backpropagation in a joint-learn fashion.

Hyper-Parameter Settings We train DCL at learning rate of 0.003, batch size at 512, training epoch at 300 and weight decay at 0.0005. Horizontal Flip is applied during training. Specifically, we set sampling scheduler to convex function in Equation 1, loss scheduler to composite function in Equation 12 with advanced self-learning point p to 0.3, and k in $\mathcal{L}_{TEA}$ (Equation 10) to 25. The margin is set to 0.2.

Time Performance We train all the models with TITAN XP GPU. Compared to the baseline DeepID2 which takes 20 hours to train, DCL training framework spends 20.5 hours to converge (only 0.5 hour more on sampling and loss calculation) under the same 300 epochs.

4.3.2 Overall Performance

We compared our proposed method DCL with DeepID2 [46], Over-Sampling and Down-Sampling in [9], Cost-Sensitive [16], Selective Learning (SL) [14], CRL[7], LMLE[19] and CLMLE[20].

Table 2 shows the overall results on CelebA. The baseline of our evaluation is the general face classification framework DeepID2[46] with standard cross entropy loss, where we achieve around 8% performance improvement. Compared to the recent advanced method, our method outperforms 3.12% to Selective Learning[14], 2.45% to CRL-I[7], 5.22% to LMLE[19] and 0.27% to CLMLE[20], respectively. Specifically, LMLE/CLMLE methods are sample-clustering based methods. However, one sample is usually bundled with multiple different attributes. It is challenging to handle all the aspects of different attributes

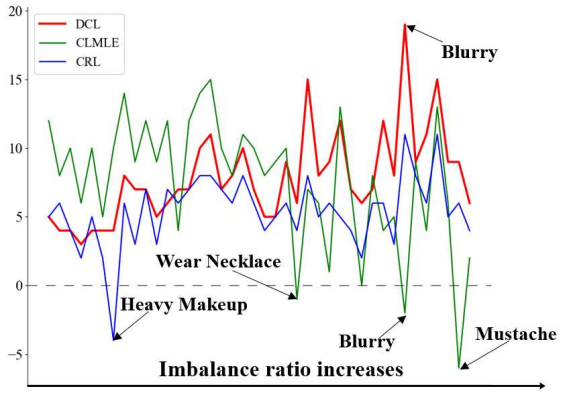

Figure 4. Comparison of performance gain to the DeepID2 for DCL, CRL and CLMLE with respect to the imbalance ratio.

Method	SS	TL	LS	Performance
1: Baseline (DeepID2)	0	0	0	81.17
2: 1 + SS	1	0	0	86.58
3: 2 + TL	1	1	0	87.55
4: 3 + LS	1	1	1	89.05

Table 3. Ablation study of each component: SS-Sampling Scheduler, TL-Triplet Loss with Easy Anchor, LS-Loss Scheduler.

Method	Performance
1: DeepID2	81.17
2: DeepID2 + Convex	**86.58**
3: DeepID2 + Linear	86.36
4: DeepID2 + Concave ($\lambda = 0.99$)	85.90
5: DeepID2 + Composite	86.07
*: DeepID2 + Linear Decreasing Imbalance	85.11

Table 4. Performance comparison between different scheduler functions selection. Method 2 in this table is corresponding to the method 2 in Table 3.

in constructing quintuplet (four-samples). In our proposed DCL method, it treats different attributes individually based on their own distributions and the triplet loss is also defined in attribute-level so that it can be easily expanded to multiple attributes learning problem. Besides, our method is computational efficient with minimal extra time cost compared to the cross-entropy loss. In LMLE/CLMLE, a computational expensive data pre-processing (including clustering and quintuplet construction) is required for each round of deep model learning. To create a quintuplet for each data sample, four cluster- and class-level searches are needed.

4.3.3 Effect of Data Imbalance Level

In this part, we show the performance gain of each attribute respecting to the data imbalance level compared with the baseline method DeepID2 in Figure 4. In the figure, red, blue, green curves indicate DCL, CRL, CLMLE respectively. The horizontal axis indicates the imbalance level and the vertical axis is the performance gain to the baseline for each method. We can observe that our proposed DCL method stably improves the performance across all the attributes while others degrade in some. Specifically, CRL is poor on attribute 'Heavy Makeup'(-4%: level-11) and CLMLE is poor on attributes 'Wear Necklace'(-1%: level-43)/'Blurry'(-2%: level-45)/'Mustache'(-6%: level-46). Our method achieves remarkable performance over the other two methods when the data is largely imbalanced, which results from the target distribution transition from imbalanced to balanced in sampling strategy. In the later stage of learning, the model focuses more on minority class while still keeps an appropriate memory for the majority class. The most significantly improved attribute is 'Blurry', with imbalance ratio 45 (8% performance gain to CRL, 21% to CLMLE). Considering all these three methods adopt the same backbone, results show the advantage of the DCL training framework.

4.3.4 Ablation Study

There are several important parts in the proposed DCL framework, including the sampling scheduler, design of the triplet loss with easy anchor and loss scheduler. We provide the ablation study in Table 3 to illustrate the advantages of each component. Sampling scheduler (SS) aims to dynamically manage the target data distribution from imbalanced to balanced (easy to hard) and the weight of each sample in $\mathcal{L}_{\text{DSL}}$ (Equation 7). Triplet loss with easy anchors (TL) modifies the anchor selection of triplet pair for better learning ($\mathcal{L}_{\text{TEA}}$). Loss scheduler (LS) controls the learning importance between $\mathcal{L}_{\text{DSL}}$ loss and $\mathcal{L}_{\text{TEA}}$ loss. From the table, we can see that our two important curriculum schedulers contribute a lot with performance gain to the whole system.

4.3.5 Effect of Scheduler Function Selection

Since we design several scheduler functions with different properties, we also include an analysis of them. The experiment setup is that we only include the selection variation for sampling scheduler, disable the metric learning with easy anchor and loss scheduler to avoid the mutual effect. In Table 4, remember that the target distribution of methods (2-5) is nonlinearly adjusted by the power operation (Eq. 6) of the scheduler function value. For method (*), the distribution is simple linearly decreasing to 1 at the end of the training. We can observe that method (*) is much worse than others. Also, the convex function is a better selection for sampling scheduler. According to the definition of scheduler function which indicates the learning speed, it interprets that it is better for the system to learn the imbalanced data slowly at the very beginning of training and then speed up for balanced data learning.

Method	Deep-Mar[30]	Inception-v2[21]	HP-net[33]	JRL[52]	VeSPA[43]	LG-Net [32]	DCL(ours)
mA	73.8	75.4	76.1	77.8	77.7	**78.7**	**83.7**

Table 5. Comaprison with the state-of-the-art methods on RAP[31] dataset. The 1st/2nd best results are highlighted in red/blue.

Imbalance Ratio (1:x)	1∼25	25∼50	>50
Baseline	79.3	68.9	68.0
DCL	83.1	83.9	85.5

Table 6. Average balanced mean accuracy (mA) in different groups of imbalance ratios. Baseline is a ResNet-50 model trained with cross entropy loss.

	Cross Entropy	CRL[7]	DCL(ours)
Accuracy	68.1	69.3 (+1.2)	71.5(+3.4)

Table 7. Results on CIFAR100 dataset (to baseline improvement).

4.4. Experiments on RAP Pedestrian Dataset

4.4.1 Implementation Details

Network Architecture We use ResNet-50[18] as the backbone for our proposed method. For each attribute, we set an extra feature layer of 64-dimension and a final output layer. Our baseline in table 6 is a ResNet-50 model trained with Cross Entropy loss in a multi-task learning framework.
Hyper-Parameter Settings We train DCL with batch size 512, learning rate 0.003, decay at 0.0005 and the epoch at 300. Horizontal Flip is applied during training. Specifically, we set sampling scheduler to convex function in Equation 1, loss scheduler to composite function in Equation 12 with advanced self-learning point p to 0.3, and k in $\mathcal{L}_{TEA}$ (Equation 10) to 25.

4.4.2 Overall Evaluation

For overall evaluation, we include several the state-of-the-art methods that been evaluated in this dataset, including Deep-Mar [30], Inception-v2 [21], HP-net [33], JRL [52], VeSPA [43] and LG-Net [32]. Table 5 indicates the average class-balanced mean accuracy (mA) for each method in RAP dataset. The 1st/2nd best results are highlighted in red/blue, respectively. We can see that our proposed DCL method outperforms the previous best one (LG-Net) with a large performance gain (5%). In term of computational complexity, methods like LG-Net and HP-net apply class-wise attention to their model, so their methods take more resource in training and inference. Our proposed method is an end-to-end framework with small extra cost.

4.4.3 Effect of Data Imbalance Ratio

Different from the definition of imbalance level (majority class rate-50%) in CelebA, imbalance ratio (1:x) in RAP is the ratio of minority samples to majority samples. As we mentioned, there are 70 attributes in this dataset and the imbalance ratio is up to 1:1800. Therefore, to show the advantage of our method for imbalanced data learning, we group attributes into three categories concerning imbalance ratio and compare the average mA with the baseline method. The baseline is a ResNet-50 model trained with cross-entropy loss. From Table 6, we can observe that for group 1 with attribute imbalance ratio from 1∼25, our method outperforms 3.8% to the baseline. When the data is more imbalance distributed in group 2 with ratio 25∼50 and group 3 with ratio >50, DCL achieves 15.0% and 17.5% performance gain, respectively. This result demonstrates that our proposed DCL method indeed works effectively for extremely imbalanced data learning.

4.5. Experiments on CIFAR-100 Dataset

To validate the generalization ability of our method, we conduct the experiment on a balanced dataset CIFAR-100 with our learning framework. In this balanced case, methods [9, 16, 14] in Table 2 are the same to the baseline method with cross-entropy loss. Also, there is no performance report of LMLE/CLMLE for generalization check. Therefore, we compare the results with the baseline and CRL[7] in Table 7. From the result, we can see our DCL method outperforms the baseline and CRL with +3.4% and +2.2%, respectively. Compared to CRL, our proposed triplet loss with easy anchor stabilizes the training process. Combined with the loss learning scheduler, DCL makes a better rectification on feature space to provide a better representation for the general classification.

5. Conclusion

In this work, a unified framework for imbalanced data learning, called Dynamic Curriculum Learning (DCL) is proposed. For the first time, we introduce the idea of curriculum learning into the system by designing two curriculum schedulers for sampling and loss backward propagation. Similar to teachers, these two schedulers dynamically manage the model to learn from imbalance to balance and easy to hard. Also, a metric learning triplet loss with easy anchor is designed for better feature embedding. We evaluate our method on two widely used attribute analysis datasets (CelebA and RAP) and achieve the new state-of-the-art performance, which demonstrates the generalization and discriminative power of our model. Particularly, DCL shows a strong ability for classification when data is largely imbalance-distributed.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM, 2009. 2

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 1, 2

[3] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004. 2

[4] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003. 2

[5] JJ Chen, C-A Tsai, H Moon, H Ahn, JJ Young, and C-H Chen. Decision threshold adjustment in class prediction. *SAR and QSAR in Environmental Research*, 17(3):337–352, 2006. 2

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. *arXiv preprint arXiv:1901.05555*, 2019. 2

[7] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. 2017. 1, 2, 5, 6, 8

[8] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1, 2, 4

[9] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003. 1, 2, 6, 8

[10] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, 20(1):18–36, 2004. 2

[11] Chen Gong, Dacheng Tao, Stephen J Maybank, Wei Liu, Guoliang Kang, and Jie Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016. 2

[12] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. *arXiv preprint arXiv:1808.01097*, 2018. 2

[13] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pages 878–887. Springer, 2005. 1, 2

[14] Emily M Hand, Carlos D Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In *AAAI*, 2018. 1, 2, 5, 6, 8

[15] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328. IEEE, 2008. 1

[16] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008. 1, 2, 6, 8

[17] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[19] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, 2016. 2, 6

[20] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Deep imbalanced learning for face recognition and attribute prediction. *arXiv preprint arXiv:1806.00194*, 2018. 2, 6

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 8

[22] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung. Classification of imbalanced data by combining the complementary neural network and smote algorithm. In *International Conference on Neural Information Processing*, pages 152–159. Springer, 2010. 2

[23] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *AAAI*, volume 2, page 6, 2015. 2

[24] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2018. 2

[25] Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Supervised neural network modeling: an empirical investigation into learning from imbalanced data with labeling errors. *IEEE Transactions on Neural Networks*, 21(5):813–830, 2010. 2

[26] Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(3):552–568, 2011. 2

[27] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 5

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[29] Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *Computer Vi-*

sion and Pattern Recognition (CVPR), 2011 IEEE Conference on, pages 1721–1728. IEEE, 2011. 2

[30] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pages 111–115. IEEE, 2015. 8

[31] Dangwei Li, Zhang Zhang, Xiaotang Chen, Haibin Ling, and Kaiqi Huang. A richly annotated dataset for pedestrian attribute recognition. *arXiv preprint arXiv:1603.07054*, 2016. 2, 5, 8

[32] Pengze Liu, Xihui Liu, Junjie Yan, and Jing Shao. Localization guided learning for pedestrian attribute recognition. *arXiv preprint arXiv:1808.09102*, 2018. 8

[33] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. *arXiv preprint arXiv:1709.09930*, 2017. 8

[34] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009. 2

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2, 5

[36] Tomasz Maciejewski and Jerzy Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*, pages 104–111. IEEE, 2011. 1, 2

[37] Krystyna Napierała, Jerzy Stefanowski, and Szymon Wilk. Learning from imbalanced data in presence of noisy and borderline examples. In *International Conference on Rough Sets and Current Trends in Computing*, pages 158–167. Springer, 2010. 2

[38] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 2

[39] Shirui Pan and Xingquan Zhu. Graph classification with imbalanced class distributions and noise. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013. 2

[40] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H Lampert. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5492–5500, 2015. 2

[41] Nikolaos Sarafianos, Theodoros Giannakopoulos, Christophoros Nikou, and Ioannis A Kakadiaris. Curriculum learning of visual attribute clusters for multi-task classification. *Pattern Recognition*, 80:94–108, 2018. 2

[42] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 680–697, 2018. 2

[43] M Saquib Sarfraz, Arne Schumann, Yan Wang, and Rainer Stiefelhagen. Deep view-sensitive pedestrian attribute inference in an end-to-end model. *arXiv preprint arXiv:1707.06089*, 2017. 8

[44] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Andres Folleco. An empirical study of the classification performance of learners on imbalanced and noisy software quality data. *Information Sciences*, 259:571–595, 2014. 2

[45] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2015. 2

[46] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014. 6

[47] James S Supancic and Deva Ramanan. Self-paced learning for long-term tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2379–2386, 2013. 2

[48] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009. 1, 2

[49] Nguyen Thai-Nghe, Zeno Gantner, and Lars Schmidt-Thieme. Cost-sensitive learning methods for imbalanced data. In *The 2010 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2010. 2

[50] Kai Ming Ting. A comparative study of cost-sensitive boosting algorithms. In *In Proceedings of the 17th International Conference on Machine Learning*. Citeseer, 2000. 1, 2

[51] Jason Van Hulse and Taghi Khoshgoftaar. Knowledge discovery from imbalanced and noisy data. *Data & Knowledge Engineering*, 68(12):1513–1542, 2009. 2

[52] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Attribute recognition by joint recurrent learning of context and correlation. 2017. 8

[53] Chan-Yun Yang, Jr-Syu Yang, and Jian-Jun Wang. Margin calibration in svm class-imbalanced learning. *Neurocomputing*, 73(1-3):397–411, 2009. 2

[54] Hualong Yu, Changyin Sun, Xibei Yang, Wankou Yang, Jifeng Shen, and Yunsong Qi. Odoc-elm: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data. *Knowledge-Based Systems*, 92:55–70, 2016. 2

[55] Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 435–442. IEEE, 2003. 1, 2

[56] Jing Zhang, Xindong Wu, and Victor S Shengs. Active learning with imbalanced multiple noisy labeling. *IEEE transactions on cybernetics*, 45(5):1095–1107, 2015. 2

[57] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006. 1, 2