Proceedings of Machine Learning Research vol 107:193–206, 2020

SelectNet: Learning to Sample from the Wild for Imbalanced Data Training

Yunru Liu E0189564@U.NUS.EDU
Department of Mathematics
National University of Singapore
10 Lower Kent Ridge Road
Singapore 119076

Tingran Gao TINGRANGAO@GALTON.UCHICAGO.EDU
Committee on Computational and Applied Mathematics
Department of Statistics
University of Chicago
5747 S Ellis Avenue Jones 316
Chicago IL 60637-1441

Haizhao Yang HAIZHAO@PURDUE.EDU
Department of Mathematics
Purdue University
150 N. University Street
West Lafayette, IN 47907-2067

Abstract

Supervised learning from training data with imbalanced class sizes, a commonly encountered scenario in real applications such as anomaly/fraud detection, has long been considered a significant challenge in machine learning. Motivated by recent progress in curriculum and self-paced learning, we propose to adopt a semi-supervised learning paradigm by training a deep neural network, referred to as SelectNet, to selectively add unlabelled data together with their predicted labels to the training dataset. Unlike existing techniques designed to tackle the difficulty in dealing with class imbalanced training data such as resampling, cost-sensitive learning, and margin-based learning, SelectNet provides an end-to-end approach for learning from important unlabelled data "in the wild" that most likely belong to the under-sampled classes in the training data, thus gradually mitigates the imbalance in the data used for training the classifier. We demonstrate the efficacy of SelectNet through extensive numerical experiments on standard datasets in computer vision.

Keywords: Imbalanced data, semi-supervised learning, classification, deep learning

1. Introduction

The success of supervised learning algorithms largely hinges upon high quality training data. Due to resource constraints and the nature of the specific applications, it can often be difficult to train a classifier on a training data set with balanced numbers of samples within each class, especially in scenarios such as anomaly detection (Hodge and Austin, 2004) and rare event discovery (Hospedales

et al., 2013). in some instances, the difference between sample sizes within classes can differ by several orders of magnitude, easily causing serious inductive bias that leads to poor prediction performance for minor classes (Dong et al., 2019; Buda et al., 2018). Unfortunately, often times in these applications it is much more important to successfully predict the minor class samples, such as disease discovery and fraud detection (Hospedales et al., 2013).

Existing techniques for tackling the data imbalance problem can be roughly divided into two categories. One is to adopt balanced strategies for the class imbalanced training data, such as bootstrapping the minor classes or downsampling the major classes, or a combination of both, usually following an ensemble learning paradigm (Chawla et al., 2002; Liu et al., 2009; Maciejewski and Stefanowski, 2011); the other is to adjust the learning objective, such as introducing different weights for samples from the major or minor classes respectively or employing a boosting strategy adapted to the heterogeneous sampling density (Zadrozny et al., 2003; He and Garcia, 2008). The two categories of methods are not mutually exclusive — in fact it is often beneficial to combine the benefits of each type of methods to achieve even better results in practice. It is worth pointing out that, however, all these techniques are crafted towards fully exploiting the structure of the skewed training data, which suffers from the deficiency in the minor training data classes.

Inspired by the emerging trend of semi-supervised learning in the past decades, in this paper we propose to borrow powers from the unlabelled data "in the wild," which are relatively easy to obtain (e.g. from modern search engines or web scrapers) but may be difficult or expensive to label (due to lack of time or human power). We hope to enlarge the training data set with more data instances of the minor classes, which balances out the skewness in the original training data, at the slight expense of incorporating few unlabelled data that are mistakenly treated as belonging to a minor class; we introduce an iterative learning strategy that is more reluctant at accepting a misclassified unlabelled data at beginning, but eventually gains more confidence and leverages the full power of unlabelled data to mitigate the imbalance issue in the original training data set. The gradual adjustment of the "attitude" towards unlabelled data is motivated by recent work in curriculum learning (Weinshall et al., 2018; Jiang et al., 2017) and self-paced learning (Kumar et al., 2010; Jiang et al., 2015); we implement this mechanism in an end-to-end fashion by means of a deep neural network, dubbed *SelectNet*. We demonstrate using extensive numerical experiments that this architecture is capable of recognizing important samples from the unlabeled data that most effectively reduces the class imbalance issue in the original training data, in such a way that the minor class prediction accuracy gets improved at no expense of sacrificing the major class prediction accuracy.

In summary, the key contributions of this paper are as follows:

- Unlike existing techniques that strive to find an appropriate way to deal with the class imbalanced issues in the original data set, we propose the novel paradigm of leveraging the unlabelled data in a semi-supervised fashion;
- We design an end-to-end deep neural network architecture, the SelectNet, which automatically learns to pick important data samples from the pool of unlabelled data and use them for improving the classifier;
- The SelectNet can be realized as an additional regularization term for a deep neural network based classifier, which can be trained along with the main classification DNN in the same computational workflow;

- Extensive numerical experiments are conducted to compare the performance of SelectNet against competing methods over standard computer vision benchmarks, and the results speak of the superior power of SelectNet in the face of severe class imbalance in the training data.

2. Related Work

2.1. Learning with Imbalanced Data

Dataset Resampling Two naive but effective way of resampling techniques are *oversampling*, which repeatedly sample data from the minor class until reaching the desired amount, and *down-sampling*, which sample the same amount of data from major class to match with minor class (He and Garcia, 2008; Chawla et al., 2002; Oquab et al., 2014). When using traditional machine learning methods like linear classifiers, oversampling can cause serious overfitting (Chawla et al., 2002). In the setting of deep neural networks, oversampling shows better compatibility while missing information in the downsampling strategy shows a critical disadvantage (Buda et al., 2018).

Cost-sensitive learning Cost-sensitive learning strategies aims that adjusting the weights in the objective loss function for training samples from different classes. Popular weight adjusting strategies include assigning weights according to inverse class frequency (Huang et al., 2016; Wang et al., 2017), or with respect to the "hardness" of the training samples, e.g., those samples that are wrongly classified by the classifier being trained (Lin et al., 2017; Dong et al., 2017). In a sense, resampling from origin class imbalanced training data set plays a similar role as assigning higher weights for the wrongly predicted samples.

2.2. Semi-supervised Learning

Self-paced Learning In (Kumar et al., 2010), the authors proposed *self-paced learning*, an iterative approach to select "easy" training samples based on the current parameters of the neural network. The number of samples selected at each iteration is gradually annealed such that in the later learning stage well-trained model can learn more samples with better tolerance to noise. In light of this, people begin to use self-paced learning and pseudo labels to refine training result, like in (Jiang et al., 2014).

A related regime is co-training (Blum and Mitchell, 1998), which alternately trains two or more classifiers, and passes "confident training samples" determined by one classifier to another classifier as training data, together with the "confidently" predicted label.

We give a slightly more detailed account of the paradigm of self-paced learning here, as this is an important motivation for our approach for tackling the imbalanced data issue. For a training data set $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_n, \mathbf{y}_n)\}$, self-paced learning uses a vector $\mathbf{v} \in \{0, 1\}^n$ to indicate whether or not each training sample should be included in the current training stage ($v_i = 1$ if the ith sample is included in the current iteration). The overall target function including $\mathbf{v}$ at iteration t is

$$(\mathbf{w}_{t+1}, \mathbf{v}_{t+1}) = \underset{\mathbf{w}\in\mathbb{R}^d, \mathbf{v}\in\{0,1\}^n}{\arg\min} \sum_{i=1}^{n} v_i \mathbb{L}(y_i, f(\mathbf{x}_i, \mathbf{w})) - \lambda \sum_{i=1}^{n} v_i \tag{1}$$

where $\mathbb{L}(y_i, f(\mathbf{x}_i, \mathbf{w}))$ denote the loss function of a convolutional neural network (CNN) model and $\mathbf{w}$ refer to the model weights. When this model is relaxed to $\mathbf{v} \in [0, 1]^n$, a straightforward derivation

easily reveals a rule for the optimal value for each entry v_i as

$$v_i = \begin{cases} 1 & \mathbb{L}(y_i, f(\mathbf{x}_i, \mathbf{w})) < \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In this formulation, the two sets of variables are model weights **w** and indicator vector **v**, and a common training strategy is to alternatively fix one set of variables and optimize the other set.

MentorNet The architecture of MentorNet was proposed in (Jiang et al., 2017) to further improve the loss thresholding strategy of self-paced learning. Instead of alternatively updating the "curriculum," the authors suggest that we use a network to directly learn the weighing curriculum (the vector **v**) from the training data, which is trained on a subset of the training data that is cleaned up and labeled with either "right" or "wrong," indicating whether the input classification label of one sample is the true label or manually disturbed one. The output vector of MentorNet is used as the weight for the loss of each training sample. MentorNet and the corresponding base net (StudentNet) are trained alternatively to provide better-labeled training data and improve the overall accuracy on noisy-labeled datasets. Inspired by its idea, we expanded our sample choosing vector into a network output, which will produce better sign of confidence than the loss value used in self-paced training.

SPARC Zhou et al. (2018) proposed the schema *SPARC* for learning network representation from rare category data, with an additional unlabeled data set. The minor groups in networks are emphasized for generating good network representations. To capture the underlying distribution of rare category examples, the predictions of both labeled and unlabeled data are considered in weighing training samples, which jointly produce separable margins between minor groups and major groups. The idea of enhancing the margin also motivated us to reuse labeled data to secure the boundary performance.

3. Algorithm

Motivated by the methodology of semi-supervised learning, the approach proposed here leverages both labeled and unlabeled data to mitigate the class imbalance issue. Intuitively, we would like to "bootstrap" the classifier by adding the unlabelled data predicted as the minor classes by the current classifier to the training set. This approach is similar toe the "pseudo-label" approach used in practice Lee (2013); Wu and Prasad (2018); the success of this procedure certainly relies on correctly identifying minor classes from the unlabelled data, and thus depends on the data distribution and the actual decision boundary. We propose to use SelectNet to learn which unlabelled data to add to the training set.

3.1. Formulation

We distinguish two sets of data, labelled and unlabelled, that are used for training. Denote $\mathcal{D} = (\mathbf{x}_i, y_i)$ for the original labeled imbalanced dataset with m classes, where y_i are one-hot encodings of the class labels. Asuume that K out of the m classes are deficient in class size and they will be referred to as the minor classes. We use C_i to indicate the ith class, and $|C_i|$ for the number of training data in this class. The ratio $\frac{\max_i |C_i|}{\min_i |C_i|}$ measures the level of imbalance of the original training set. In addition, we assume an extra "pool" of unlabelled data $\mathcal{U} = (\mathbf{x}_i)$ is available, from which we collecte more training samples in the minor classes using the current classifier. In practice, these

unlabelled data may be collected by keywords searching in google or crawled from the internet. The hypothesis on $\mathcal{U}$ is that they have high potential of including data in minor classes.
For a classification task, denote its loss function as $\mathbf{L}_c(y_i, f_c(\mathbf{x}_i, \mathbf{w}_c))$, where $f_c(\mathbf{x}_i, \mathbf{w}_c)$ is the main deep neural network with weights $\mathbf{w}_c$ being trained for the classification task. In each iteration, the label of an unlabelled sample $\mathbf{x}_i \in \mathcal{U}$ is inferred from the main classifier and is denoted as $\hat{y}_i = argmax(f_c(\mathbf{x}_i, \mathbf{w}_c))$. In the meanwhile, another deep neural network is trained to determine which of the unlabelled data will be added to the training sample for the next iteration. The second neural network is denoted as $f_s(\mathbf{z}_i; \Theta)$ with parameters Θ. $f_s(\mathbf{z}_i; \Theta)$ takes as input the last layer feature of an unlabeled data output by f_c along with its loss calculated by the classification model, i.e., $z_i = (f_c(\mathbf{x}_i, \mathbf{w}_c), \mathbf{L}\left(f_c(\mathbf{x}_i, \mathbf{w}_c), \mathbf{w}_c\right))$. The entire model has the following form:

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^d, f_s\in\{0,1\}^{n\times m}} \mathbb{F}(\mathbf{w}, \mathbf{v}) = & \frac{1}{n_D} \sum_{i\in\mathcal{D}} \mathbf{L}_c(y_i, f_c(\mathbf{x}_i, \mathbf{w}_c)) \\ & + \frac{1}{n_{add}} \sum_{i\in\mathcal{D}'\cup\mathcal{U}'} \left[f_s(\mathbf{z}_i; \Theta)\mathbf{L}_c(\hat{y}_i, f_c(\mathbf{x}_i, \mathbf{w}_c)) - \lambda f_s(\mathbf{z}_i; \Theta)\right], \end{aligned} \tag{3}$$

where $\mathcal{D}'$ and $\mathcal{U}'$ include the data whose top-1 prediction result is a minor class. The training procedure alternatively updates the main classification network $f_c(\mathbf{x}_i, \mathbf{w}_c)$ and the selection network $f_s(\mathbf{z}_i; \Theta)$. We remark that the "selection" part includes an additional regularization term with a penalty parameter λ, since f_s will trivially attain zero value at all unlabelled data if $\lambda = 0$. Moreover, only samples which are confidently predicted in a minor class are added to the training set so as to prevent the model from further aggravating data imbalance. Evaluation of the confidence differs from method to method. For example, in self-paced learning, loss value is directly used as a sign of confidence. The samples with loss value smaller than a hyperparameter λ will be selected. We will introduce our indicator of confidence Section 3.3.

3.2. Context Data

When training from class-imbalanced data, the accuracy of the classifier is mainly hampered by its weak performance on minor classes. There are two sources of this inaccuracy, as schematically illustrated in Figure 1: the low recall — the situation that samples in minor classes are classified as a major class, or the low precision, where samples from a major class are classified as a minor class. Empirically, we observe when training on the original imbalanced dataset that it is the low recall that affects the prediction accuracy, whereas the precision was relatively high, which is as expected from a small class size. As more unlabelled minor class data are added to the training set, the classifier gradually gains higher recall, but the precision drops as well, which we conjectured is due to the quality of added data, especially those that are added with wrongly predicted labels. This situation is similar to the "noisy label" setting described in Ma et al. (2018), where the training data contains incorrect labels. To improve the recall while keeping the precision from dropping, we also identify samples from the labelled dataset that are wrongly predicted as a minor class, and add them to the training data; these samples are "hard" to classify, which suggests they play important roles in characterizing the decision boundary efficiently.
Therefore, in each iteration step of the proposed schema (3), the selected data to be added to training consist of two parts: the unlabeled data from $\mathcal{U}$ and wrongly labeled data from $\mathcal{D}$. We refer to the second part as *context data* as it helps differentiating minor class samples from their neighboring major class samples. The label of each selected sample from the unlabelled data (in the second part

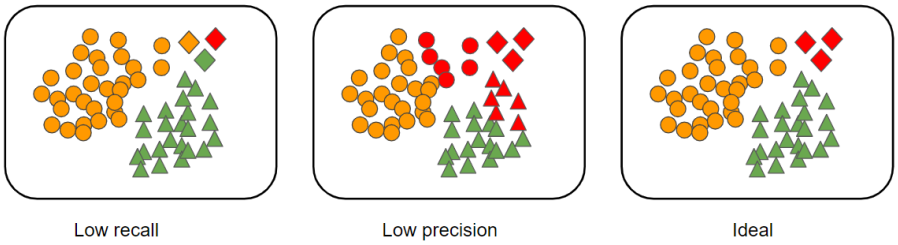

Figure 1: Decision boundary in different situation

of Formula (3) $\hat{y}_i$) is given by the prediction of $f_c(\mathbf{x}_i, \mathbf{w}_c)$, while the label for each sample selected from labelled data is used in both the first term and the second term of Formula (3).

3.3. SelectNet

Ideally, the minor class samples in $\mathcal{U}$ should be correctly identified, which will maximize the effect of using extra data. As this rarely occurs in practice, weighing the gain of collecting more information about the minor class samples and the loss of training with noisy labels is critical. This is trade-off is reflected in the second line of Formula (3), where f_s are indicators of whether a sample should be added into loss calculation or not, and the parameter λ balances out the "gain" and the conceptual "loss." In previous works such as self-paced training, f_s represents the decision made according to a predefined choosing rule, often resulted from comparison between the loss value and a threshold λ, which is exactly the derivative result of Formula 1 with respect to the hyperparameter λ:

$$f_s = \begin{cases} 1 & \text{if } \mathbf{L}\left(y_i, f_c\left(\mathbf{x}_i, \mathbf{w}_c\right)\right) < \lambda \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Such a deterministic rule for determining f_s imposes a dependence on the correct choice of hyperparameter λ, which limits the flexibility of the proposed method. We thus propose a data-driven way, using a regression model to produce f_s for each sample, motivated by the architecture of MentorNet (Jiang et al., 2017) which proposed similar idea for resolving the noisy label issue in supervised learning. In their setting, the extra network is trained on a predefined small set of samples for deciding whether a predicted label is correct or not, which can not be directly used for solving the class imbalanced issue in our problem. The intuition of turning f_s into a data-driven indicator is more on directly expanding the search space of overall target function (3).
Based on these considerations, we revise the target objective function for training f_s as

$$\hat{f}_{si} := \arg\min_{\Theta} \frac{1}{n_{add}} \sum_{i \in \mathcal{D}'} \hat{f}_s(\mathbf{z}_i; \Theta)(\mathbf{L}_{ci} - \lambda), \tag{5}$$

and f_s is assigned with respect to

$$f_{si} = \begin{cases} 1 & \hat{f}_{si} > \lambda \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

As f_s is trained to minimize the formulation in (5), the choice of λ is insensitive in a range, for the update of f_s will try to fit to it.

The purpose of training in this step is to have the second network produce accurate "approximate loss value" given a classification output for an unlabeled data point, which will be used by the main network to estimate the confidence of using the output as fake label.
In the experiment we train f_s using a simple two-fully-connected-layer network. No groundtruth label is used in training f_s. Instead, the input is the classification output from the main network, and training target is the loss of the main network incurred at the particular training sample.
Our final algorithm, the SelectNet, incorporates the ideas of using context data, and alternatively updates the selection part $\hat{f}_s$ and the base part of the model on the selected samples. The full algorithm is described in Algorithm 1.

Algorithm 1 SelectNet

Input: labeled dataset $\mathcal{D}$, unlabeled dataset $\mathcal{U}$, minor classes $\mathcal{C}$
Output: classifier with weights $\mathbf{w}$
Initialize model weight $\mathbf{w}^0$ by training on $\mathcal{D}$ with oversampling
Update w by repeat below routine t times:
1. $\mathcal{D}'$=$\{$prediction($\mathcal{D} \cup \mathcal{U}$, $\mathbf{w}^{t-1}$) in $\mathcal{C}\}$
2. Update $\hat{f}_s^t$ by $\{f_c(\mathcal{D}', \mathbf{w}^{t-1}), L_c(\mathcal{D}', \mathbf{w}^{t-1})\}$, as formula (5)
3. $f_s^t = \mathbf{1}_{(prediction(\mathcal{D}', \hat{f}_s^t) > \beta)}$, as formula (6)
4. Decide new data to be added in this routine $\mathcal{D}_{add}$ = $\{\mathcal{D}'$ when ${f_s}^t = 1\}$
5. Combine the labeled data and new data to be the training data of this routine $\mathcal{D}_{train}^t = \mathcal{D} \cup \mathcal{D}_{add}$
6. Train and update $\mathbf{w}^{t-1}$ to $\mathbf{w}^t$ using $\mathcal{D}_{train}^t$

4. Numerical Experiments

We present numerical results to support the proposed SelectNet and compare it with baselines and state-of-the-art algorithms, e.g. the imbalanced training results, methods based on oversampling (Chawla et al., 2002; Kubat et al., 1997), the application of self-paced method in imbalanced problem (Jiang et al., 2015; Zhou et al., 2018), methods based on context data (Zhou et al., 2018), and also a class-balancing loss method(Cui et al., 2019) as summarized in Section 4.2.

4.1. Datasets

CIFAR-10 and CIFAR-100 datasets (Krizhevsky and Hinton, 2009) are adopted throughout this paper. Note that CIFAR-10 and CIFAR-100 contain equal amounts of training samples in each class. Hence, the classification accuracy of the original balanced datasets serves as the upper bound of the achievable classification accuracy.
We artificially create imbalanced datasets using CIFAR-10 and CIFAR-100. In CIFAR-10 experiments, after selecting the minor classes, e.g., classes [0,2,6,7] in our experiments, we create imbalanced datasets with an imbalanced ratio 90. In particular, 1% of training samples in the selected minor classes are kept as labeled data and the other 99% of data in the selected monirity classes are left as unlabeled data. Similarly, we select clases [10,20,60,70] and [5,10,11,18,30,45,55,79,86,98] as the minor clases. Then, 90% of the data belonging to a major class are kept as labeled data and the other 10% of data in the major class are put into the unlabeled dataset. In CIFAR-100 experiments, we create imbalanced datasets with an imbalanced ratio 14 by keeping 5% of training samples in minor classes and 90% of training samples in major classes.

Recently, a long-tailed CIFAR dataset was proposed in (Cui et al., 2019) and a new method for imbalanced dataset based on a class-balanced-loss was also proposed in (Cui et al., 2019). Hence, we compare SelectNet with the class-balanced-loss method using the long-tailed CIFAR dataset. In this experiment, all the unused training samples are collected as unlabeled samples.

4.2. Methods for Comparison

Imbalanced Training In this method, the classifier is only trained with the labeled imbalanced dataset.

Oversampling The oversampling method aims at creating a balanced training dataset using the original imbalanced dataset. Suppose a certain class is minor, it repeatedly samples with replacement from the data in this class and put them together untill the number of samples is as large as that of a major class. Then these new samples are added to augment the training data. Such a process is repeated to eliminate minor classes to achieve a balanced training dataset. Finally, a classifier is trained with this new balanced dataset.

Self-paced training The idea of self-paced training can be simply migrated to the data imbalance problem. Suppose an unlabeled dataset is available for the self-paced training. After every n epochs of iterations, a certain amount of data is selected from the unlabeled dataset to augment the labeled dataset with labels given by the current classifier. The augmented dataset serves as the new training data for the next n epochs. The selection process is controlled by a threshold λ, which is set as 0.6 in this paper. When an unlabeled sample is predicted as the minor class with a loss smaller than λ using the current classifier, this sample will be added to the training dataset.

Context data As described in Section 3.2, in addition to the selection of unlabeled data as in the self-paced training, labeled data which are classified as minor classes by the current classifier can also be added to the training dataset for latter training. This approach is named as the context data method. Similar to the self-paced method, a threshold 0.6 is set to control the selection process. When a labeled sample is predicted as the minor class with a loss smaller than λ, this sample will be duplicated and added to the training dataset.

SelectNet As described in Section 3.3, after every n epochs, a deep neural network is trained to decide whether a sample should be added in the next stage of training. The model we use is a fully-connected ReLU neural network with two hidden layers of width 8 and 4, respectively. The output of this network is a 1-d scalar from a Sigmoid activation. If the output of the network is larger than a hyperparameter $\lambda = 0.6$, the corresponding input sample will be added to the training dataset for latter iterations.
For the first two methods, we set the number of epochs to be 200, while for the last three methods, we update the training samples every 10 epochs for 20 iterations, i.e., the total number of epochs is also 200.

4.3. Comparison on CIFAR imbalanced dataset

In this experiment, we show the universal advantage of the proposed SelectNet over various deep learning methods like the imbalanced train method, the oversampling method, the self-paced method, and the context data method. To make consistent comparisons, we fix the same kind of deep neural

Table 1: CIFAR accuracy comparison

	CIFAR-10			CIFAR-100		
		4 minors		4 minors		10 minors
Method	Simple net	ResNet20	ResNet56	ResNet20	ResNet56	ResNet56
Imbalanced train	0.5617	0.5659	0.5672	0.5729	0.5806	0.589
Oversampling	0.5744	0.6629	0.6897	0.5749	0.6128	0.6014
Self-paced	0.6933	0.7894	0.7951	**0.5971**	0.5901	0.5974
Context data	0.7403	0.7881	0.7864	0.5889	0.6285	0.6165
SelectNet	**0.7424**	**0.7895**	**0.8011**	0.5921	**0.6290**	**0.6171**

Table 2: CIFAR-10: class-wise prediction accuracy.

	0	1	**2**	3	4	5	**6**	**7**	8	9
Oversampling	**0.05**	0.85	**0.02**	0.41	0.49	0.50	**0.06**	**0.09**	0.68	0.79
Self-paced	**0.71**	0.90	**0.20**	0.52	0.71	0.64	**0.52**	**0.72**	0.86	0.86
SelectNet	**0.74**	0.89	**0.54**	0.57	0.71	0.69	**0.72**	**0.76**	0.86	0.85

network to implement different classification methods above. To see the influence of different network architectures on the performance of the classification methods, commonly used architectures have been explored in the test, e.g., the standard network example in Keras (Chollet et al., 2015), ResNet-20, and ResNet-56.

Table 1 summarizes the experiment results for CIFAR imbalanced datasets. Numerical results show that SelectNet is almost consisently better than other methods for both CIFAR-10 and CIFAR-100 datasets and various network structures. There is only one case in which SelectNet is the second best method with an accuracy only slightly smaller than that of the self-paced method, when the network is ResNet20 and the dataset is CIFAR-100.

Note that SelectNet is build on top of the context data method and an extra deep neural network for adaptively update training data. Hence, the results of the SelectNet is always better than the context data method, the results of which is already very promising. The consistent advantage of the SelectNet over the context data method validates the proposal of an extra deep neural network for updating training data adaptively.

Table 2 shows the category-wise f1-score of three methods for CIFAR-10 in the simple net experiment. The bold numbers are the performance of the minor classes. Each of these suffers poor accuracy when using the oversampling method. The latter two methods, with the help of additional data, considerably increased the performance of minor classes. Meanwhile, our propose SelectNet exhibits the best improvements.

Table 1 compares various methods in terms of the overall classification accuracy for all classes, while Table 2 compares these methods in terms of the classification accuracy within individual classes. It is worth emphasizing that the SelectNet significantly outperforms other methods in the classification of minor classes, which is of special interest in real applications, especially in medical applications where minor classes are more valuable.

To check the number of extra training samples added during the training of SelectNet, we choose

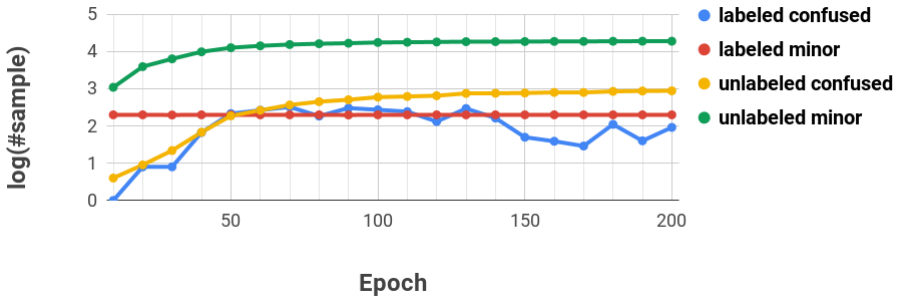

Figure 2: Number of data chosen during training.

Table 3: Long-Tailed CIFAR accuracy comparison.

Method	CIFAR-10-0.01	CIFAR-100-0.01
Class-Balanced Loss	0.7457	0.3960
SelectNet	**0.7576**	**0.4056**

the case when networks are carried out via the simple one in Keras (Chollet et al., 2015) and visualize the numbers in Figure 2. The "labeled-confused" number represents the number of labeled samples that are wrongly predicted (no matter major or minor) and added to the training dataset. The "labeled-minor" number denotes the number of labeled samples that are correctly classified as a minor sample. Similarly, the "unlabeled-confused" and "unlabeled-minor" numbers mean the numbers in the case of unlabeled samples. These numbers seem to be bounded by a constant number, which means that the mistakes SelectNet makes would not increase after a certain number of epochs. The "unlabeled-minor" number is significantly larger than other numbers, indicating the effectness of the SelectNet since most of the selections it makes are correct.

4.4. Comparison on long-tailed CIFAR dataset

Here, we compare SelectNet with the newly proposed class-balanced-loss method in (Cui et al., 2019) using the long-tailed CIFAR dataset therein. Since there is no class that has a number of samples significantly smaller than others, we treat half classes with a smaller amount of data as the minor classes. In particular, in the long-tailed CIFAR-10 dataset, the minor classes are 5,6,7,8,9; while in the case of long-tailed CIFAR-100, the minor classes are 50 to 99. All the other settings remain the same as in previous experiments. The comparison is conducted with ResNet32 and the results are summarized in Table 4. The proposed SelectNet outperforms the class-balanced-loss method by 1% classification accuracy. Which shows a universal efficacy of our method.

4.5. The choice of λ

The algorithm is designed to let the overall model fit to the hyperparameter λ, therefore the value of λ is insensitive in a range. Our experiments shows choosing λ in [0.5, 0.7] will not affect the performance.

Table 4: λ schema comparison.

λ	CIFAR-10, 4 minors, Simple Net
fixed, 0.6	0.7424
0.4 to 0.6 in 20 iterations	0.7328
0.6 to 0.4 in 20 iterations	0.7109

We also considered an schema to moving the value of λ during training. The result are worse, as shown in 4. Which may indicate that the whole process is indeed fitting to the hyperparameter as we expected, so the change of this parameter will not work.

5. Conclusion

In this work, we drew an analogy between the imbalanced data problem and semi-supervised learning, and proposed a simple yet powerful approach, referred to as SelectNet, to mitigate the class imbalance issue using unlabeled data "in the wild." We began with the observation that incorporating "context data" into training significantly improves the classification performance, then generalized the 0-1 selection rule to a continuously valued regression network that takes real values between 0 and 1 as "selection." Combining the idea of context data and minor class data selection provides significant improvement of the classification performance over existing works. We expect other cost-sensitive learning techniques to benefit from this "data side improvement" as well; we will explore this in a future work. Besides, how distribution of unlabeled data affect the classifier as well as the possible bias introduced to the classifier, are also interesting topics that can be further explored.

Acknowledgments

H. Y. was partially supported by National Science Foundation under award DMS-1945029.

References

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM, 1998.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

François Chollet et al. Keras. https://keras.io, 2015.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. *CoRR*, abs/1901.05555, 2019. URL http://arxiv.org/abs/1901.05555.

Q. Dong, S. Gong, and X. Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1367–1381, June 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2832629.

Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1851–1860, 2017.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, 21(9):1263–1284, 2008.

Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, Oct 2004. ISSN 1573-7462. doi: 10.1023/B:AIRE.0000045502.10941.a9. URL https://doi.org/10.1023/B:AIRE.0000045502.10941.a9.

T. M. Hospedales, S. Gong, and T. Xiang. Finding rare classes: Active learning with generative and discriminative models. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):374–386, Feb 2013. ISSN 1041-4347. doi: 10.1109/TKDE.2011.231.

Timothy M Hospedales, Shaogang Gong, and Tao Xiang. Finding rare classes: Active learning with generative and discriminative models. *IEEE transactions on knowledge and data engineering*, 25 (2):374–386, 2013.

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.

Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 547–556, 2014.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. Self-paced curriculum learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2694–2700. AAAI Press, 2015. ISBN 0-262-51129-0. URL http://dl.acm.org/citation.cfm?id=2886521.2886696.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.

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

Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186. Nashville, USA, 1997.

M. P. Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1189–1197. Curran Associates, Inc., 2010.

Dong-hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, 2013. Workshop on Challenges in Representation Learning, ICML, vol. 3, p. 2. 2013.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

X. Liu, J. Wu, and Z. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, April 2009. ISSN 1083-4419. doi: 10.1109/TSMCB.2008.2007853.

Xingjun Ma, Yisen Wang, Michael E. Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3355–3364, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

T. Maciejewski and J. Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 104–111, April 2011. doi: 10.1109/CIDM.2011.5949434.

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.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5238–5246, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

H. Wu and S. Prasad. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259–1270, March 2018. ISSN 1057-7149. doi: 10.1109/TIP.2017.2772836.

B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pages 435–442, Nov 2003. doi: 10.1109/ICDM.2003.1250950.

Dawei Zhou, Jingrui He, Hongxia Yang, and Wei Fan. Sparc: Self-paced network representation for few-shot rare category characterization. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2807–2816. ACM, 2018.