LEARNING TO BALANCE WITH INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Classification tasks require balanced distribution of data in order to ensure the learner to be trained to generalize over all classes. In realistic settings, however, the number of instances vary substantially among classes. This typically leads to a learner that promotes bias towards the majority group due to its dominating property. Therefore, methods to handle imbalanced data is crucial for alleviating distributional skews and fully utilizing the under-represented data. We propose a novel training method, Sequential Targeting, that forces an incremental learning setting by splitting the data into mutually exclusive subset and adaptively balancing the data distribution as tasks develop. To address problems that arise within incremental learning, we apply dropout and elastic weight consolidation with our method. It is demonstrated in a variety of experiments on both text and image dataset (IMDB, CIFAR-10, MNIST) and has proven its superiority over traditional methods such as oversampling and under-sampling.

1 INTRODUCTION

In a real-world application setting, it is rarely the case where the discrete distribution of the data is perfectly balanced. It is common for real-world data to be skewed to majority classes when the minority class is often the class of interest. Skewed data distribution is explicitly addressed in anomaly detection as it is a matter of high interest in various fields including disease, fraud and malware detection[Rao et al. (2006); Wei (2013); Cieslak et al. (2006)]. Since correctly classifying under-represented classes is equally important as classifying majority classes, if not more so, methods to temper the model from biasing towards certain classes are of great importance. In order to develop an intelligent classifier, approaches to generalize over different distributional tasks are getting increasingly important and being extensively researched.

Learning from imbalanced data inevitably brings bias toward frequently observed classes. Data-level manipulation tries to under-sample majority classes or over-sample minority classes. But these methods have a tendency to discard valuable information from observations of majority classes or over-fit to limited data, especially as the imbalance level gets severe.

We propose a novel training architecture, **Sequential Targeting (ST)**, which handles the data imbalance problem without manipulating the data by forcing an incremental learning setting. ST divides the entire training data set into mutually exclusive partitions, adaptively balancing the data distribution among tasks. The split data is then sorted in similarity with the target distribution and used to train a balanced learner. In order to address *catastrophic forgetting* (French, 1999), which is an inevitable phenomenon when training a learner in a sequential fashion, we utilize different methods[Kirkpatrick et al. (2016); Goodfellow et al. (2013)] to stabilize the knowledge attained from the previous tasks. This allows the learner to pay more focus on the newly adapted data distribution while not forgetting the representation from previous tasks.

We validate the effectiveness of our method with experiments conducted on both text and image data. Different data imbalance levels as well as proportion of the minority classes were experimented. Experimental results show that Sequential Targeting outperforms previous approaches, with notable gap especially in extremely imbalanced cases. Furthermore, Sequential Targeting proves to be compatible with both traditional data-level methods and recent algorithm-level approaches.

Our contribution in this paper is as follows:

- We introduce a novel method addressing the data imbalance problem by splitting the data and incrementally training a learner to perform a balanced learning by paying equal attention to the minority data as to the majority. The learner's initial focus on the majority class is compensated by the redistributed task that comes sequentially, resulting in an overall increase of performance.
- Incremental learning techniques have been utilized to prevent catastrophic forgetting on tasks encountered before. The novelty of our method stands since we applied incremental learning to address the data imbalance problem.
- We propose a novel method that is compatible with previous methods addressing the data imbalance problem. Best performance is shown when utilized together.

2 Related Works

Balancing Methods Previous researchers have proposed data-level methods as well as algorithm-level methods that address the data imbalance problem. Data-level techniques such as SMOTE and its variants [Chawla et al. (2002); Han et al. (2005)] attempt to alleviate the level of imbalance by manipulating the training data. Cluster based oversampling (Puntumapon et al., 2016) attempts to cluster the data into groups using k-means algorithm and apply oversampling to each cluster, which proves to reduce within-class imbalance and between-class imbalance. SMOTEboosting (Chawla et al., 2003) applies SMOTE before every round of boosting and provides more samples of misclassified set for weak learners often including the minority class samples.

Alternatively, algorithm-level methods[Krawczyk (2016); Sheng & Ling (2006)], commonly implemented with a weight or cost schema, modify the underlying learner or its output to reduce bias towards the majority group. Algorithm-level methods modify the structure of the decision process of how much to focus on under-represented samples. This could be implemented by assigning cost matrix as a penalty (Fernando et al., 2017). Moreover, loss functions could be modified, such as in the case of focal loss (Lin et al., 2017), which reshapes the standard cross entropy loss such that it penalizes the loss assigned to well-classified instances.

Catastrophic Forgetting Researchers have investigated methods to circumvent *catastrophic forgetting*, a phenomenon of a learner forgetting about the previously learned task when encountered with another task. One of the major approaches is to use an ensemble of networks, each trained with individual tasks (Rusu et al., 2016). However, this approach has a complexity issue. Alternatively, Fernando et al. (2017) proposed an ensemble approach which attempts to fix the parameters learned from the previous task and train new parameters on consecutive tasks. This method has successfully reduced complexity issues, but performance suffers from the lack of trainable parameters. Goodfellow et al. (2013), on the other hand, investigated catastrophic forgetting in neural networks and showed that applying dropout substantially helps overcome this phenomenon without any apparent downfalls. This accounts to the increased optimal size of the parameters with dropout applied, thereby increasing the capacity of the learner and reducing generalization error.

Recent studies have focused on developing a regularization term that buffers the model from forgetting the previously trained information. Elastic weight consolidation (EWC) (Kirkpatrick et al., 2016) is one of the most conspicuous works in this field. EWC algorithm, which implements modified regularization term that consolidates knowledge across tasks, imposes restriction on the model to slow down updating certain important parameters from the previous task.

EWC considers neural network training from a probabilistic perspective. Optimizing the network parameters θ is equivalent to finding their most feasible values given some data D. We can compute this conditional probability $p(\theta|D)$ from the prior probability of the parameters $p(\theta)$ and the probability of the data $p(D|\theta)$ by using Bayes' rule:

$$\log p(\theta|D) = \log p(D|\theta) + \log p(\theta) - \log p(D) \tag{1}$$

By assuming that the data is split into two independent parts, one defining the previous task D_{prev} and the other current task D_{curr}. we can rearrange the equation 1 by applying Bayesian update rule:

$$\log p(\theta|D) = \log p(D_{curr}|\theta) + \log p(\theta|D_{prev}) - \log p(D_{curr}) \tag{2}$$

The prior distribution learned from previous dataset is further enriched by the data given at the current task. Due to the intractability of the posterior distribution, EWC poses an assumption on the prior distribution to follow a Gaussian distribution with the mean as θ^*_{prev} and the Fisher information matrix, F, of the previous task as the precision. In order to minimize the loss of information, the objective function is defined as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_{curr}(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{prev,i})^2 \tag{3}$$

where $\mathcal{L}_{curr}(\theta)$ sets the loss for the current task, λ sets the importance of the previous task and i labels each parameter.

3 SEQUENTIAL TARGETING

We first introduce a broad overview of the novel training architecture: Sequential Targeting. Next, we show how this method has been applied to address the data imbalance problem.

3.1 MATCHING THE TARGET DISTRIBUTION

We propose a novel model architecture of **forced incremental learning** on imbalanced setting. The term *forced incremental learning* has rarely, if ever, been used since incremental learning is a much complicated task. In incremental learning, the new data is referred to as different *task* in this paper. The task is consistently provided therefore the learner needs to be constantly updated on the new task. Because of *catastrophic forgetting*, learners perform much better when trained with an individual task than continually being updated with multiple tasks. In spite of this phenomenon, we have proven it is beneficial to force an incremental learning setting where the given data distribution varies substantially from the target data distribution, such as in the case of data imbalance. The **target distribution**, denoted as $\mathcal{P}_T$, is the idealistic data distribution in which the learner would perform the best if trained on.

Our method effectively improves the model by dividing a given task, $\mathcal{D}_{total}$, into multiple tasks, $\mathcal{D}_1, \mathcal{D}_2, ...\mathcal{D}_k$, so that $\bigcup \mathcal{D}_i = \mathcal{D}_{total}$ and $\bigcap \mathcal{D}_i = \emptyset$. Each task is partitioned into varying distributions: $\mathcal{P}_1, \mathcal{P}_2, ...\mathcal{P}_k$. The learner is sequentially trained on these tasks in the order of similarity with the target distribution, which is measured with KL-divergence. The following has to hold:

$$D_{\mathrm{KL}}(P_T \| P_{prev}) > D_{\mathrm{KL}}(P_T \| P_{curr}), \quad where\ P_i \xrightarrow{\mathcal{D}} P_T \tag{4}$$

Using a single learner over all tasks, we incrementally condition on the maximum performance from the previous task and stabilize the learned parameters on the current task. As explained in equation 2, the learner's parameters from the previous task distribution is used to train on the current task distribution. KL divergence explains the discrepancy between the task distribution and the target distribution. As tasks proceed, the data distribution approximates the target distribution. At last, the task becomes identical to the target distribution ensuring $D_{\mathrm{KL}}(P_T \| P_k) \approx 0$.

The number of data splits, k, and how each task is partitioned to have varying KL-divergence values are both highly dependent on what $\mathcal{P}_T$ is defined to be. We believe this training approach is the first of its kind, with the best of our knowledge.

3.2 ADAPTIVE BALANCING FOR DATA IMBALANCE

Imbalanced data setting tempers the model from learning extensively from under-represented instances. Therefore, it is crucial to enforce the learner to acquire knowledge equally among classes. In an idealistic setting, a learner is trained with equally represented dataset. However, in a realistic setting, imbalanced data includes some under-represented classes and the learner has difficulty

acquiring sufficient knowledge to generalize. Sequential Targeting enables balanced learning by redistributing the task to approach the target distribution as tasks develop.

It is idealistic to assume $\mathcal{P}_T \overset{d}{=} Uniform[0, p]$ in an imbalanced data setting therefore when p denotes the number of classes the target distribution is discrete uniform: $\{\frac{1}{p}, \frac{1}{p}, , , , , \frac{1}{p}\}$. In our method, the training data is redistributed into two splits so that the last split is identical to the uniform distribution.

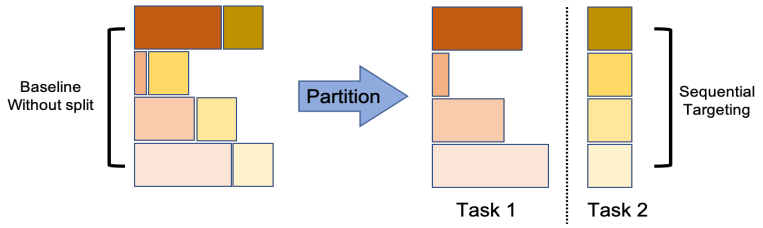

Figure 1: Comparison between baseline and Sequential Targeting

This training architecture lets the learner pay more focus to the under-represented data by manipulating the learning sequence. Sequentially training the learner to be exposed to increasing portion of minority class data benefits the overall performance. Moreover, applying dropout to the layers and implementing EWC during the transfer between tasks proves to help the learner maintain the knowledge acquired from the previous split.

4 Experiments

4.1 Evaluation Metrics and Ratios

Accuracy is commonly used to measure the performance of a classification model. However, when it comes to skewed data, accuracy alone can be misleading and thus other appropriate metrics are needed to correctly evaluate the performance of the model. In this paper, we use precision, recall, and macro F1-score to objectively evaluate the model in a skewed data setting. *Precision* measures the percentage of actual positive among the number of positively predicted samples. *Recall* measures the percentage of the truly positive instances that was correctly predicted by the model. As precision and recall is in a trade-off relationship, selecting a learner that performs well on both metrics would be a reasonable policy. *Macro F1-score* combines both precision and recall as a harmonic mean weighted with equal importance on each class rather it be sparse or rich. In this paper, F1-score is used as the core metric for measuring performance.

Following the conventions of the previous research on imbalanced data[Buda et al. (2018); Johnson (2019)], we employ three distinct ratios used throughout the experiment to represent the imbalanced state of the data. One is the proportion of minority classes over majority classes μ:

$$\mu = \frac{|\{i \in 0, 1, \ldots, N : C_i \text{ is minority}\}|}{N} \tag{5}$$

where C_i is the set of instances in class *i* and *N* is the total number of classes. Another parameter ρ is a ratio between the number of instances in majority classes and the number of instances in minority classes defined as follow:

$$\rho = \frac{\{\max_i(C_i)\}}{\{\min_i(C_i)\}} \tag{6}$$

The other parameter η is a parameter that compares the relative number of minority class instances among splits. For instance, if the first task consists of 100 samples in minority class and the second task consists of 50 samples, then η will be 2:1. We experienced various combinations of these three ratios and concluded that η does not have a significant effect on model performance. Therefore, the

number of instances of the minority classes between the splits is fixed as 1:1 throughout the whole experiment. Further consideration of η can be found in the discussion section.

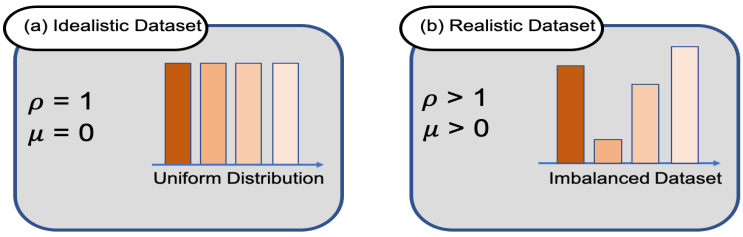

Figure 2: Idealistic and Realistic data distributions

Figure 2 shows variation of μ and ρ in the case of a idealist dataset and a realistic dataset.

4.2 Dataset and neural network architecture

In this research, we performed experiments on IMDB, CIFAR-10, and MNIST datasets. The datasets were deliberately made into varying imbalanced states as shown in Table 1.

Table 1: Simulated Dataset in variations of ρ and μ.

Dataset	ρ	μ	Class	Train		Validation		Test	
				Minority	Majority	Minority	Majority	Minority	Majority
IMDB	10	N/A	2	1,250	12,500	2,500	2,500	10,000	10,000
	20			625	12,500	2,500	2,500	10,000	10,000
	50			250	12,500	2,500	2,500	10,000	10,000
CIFAR-10	10	0.8	10	400	4,000	100	100	100	100
	5	0.6		800	4,000	200	200	200	200
MNIST	10	0.8	10	500	5,000	100	100	100	100

IMDB is a text dataset, which contains 50000 movie reviews with binary labels (positive/negative). Reviews have been preprocessed, and each review is encoded as a sequence of word indexes. Three different imbalance ratios have been deliberately made ($\rho = 10, 20, 50$) to test how each method performs as the imbalance level worsens. The positive reviews are regarded as the positive class in our experiment. A CNN + LSTM model architecture was used for this task. After an initial embedding layer, a 1-dimension convolution layer is followed with a dropout layer ($dr = 0.2$). It is then followed by a Bidirectional LSTM and a Uni LSTM layer. Lastly, a 1-dimensional fully-connected layer ($dr = 0.2$) is followed by a sigmoid activation.

CIFAR-10 is a image dataset that consists of 10 classes with 6000 training and 1000 test data for each class. We sampled from the original data with predetermined ρ ratio ($\rho = 5, 10$) in order to experiement on an imbalanced setting. Since there are multiple classes in this dataset, two varied fraction of minority classes ($\mu = 0.8, 0.6$) were simulated and tested as well. Following the work of Masko & Hensman (2015) on CIFAR-10, we used a variant of CNN (Lecun et al., 1998). After applying two dimensional convolutional layer, ReLU activation is applied followed by maxpooling and a dropout layer ($dr = 0.5$). This procedure is repeated twice and three fully-connected layers with hidden node size of 120, 84, and 10 are utilized.

MNIST is a image dataset that consists of simple handwritten single digits. Each class of digits consists of 5500 training, 500 validation and 1000 test data. MNIST data is deliberately manipulated in order to match the predetermined ρ and μ ratio($\rho = 10$, $\mu = 0.8$). The model architecture used in MNIST data is a simple Multi-Layer Perception network which includes three consecutive fully-connected layers with hidden node size of 512, 512, and 10. A dropout layer with dropout rate 0.2 is implemented on each hidden layer. The activation function used in this architecture is ReLU.

Experimental setup. In our experiments, our proposed method has been extensively compared with two data-level methods, random oversampling (ROS) and random under-sampling (RUS). A naive duplicate sampling approach has been used for oversampling the minority data for simplicity. We

explored the full capability of Sequential Targeting by comparing multiple combinations with ROS and EWC. ROS is combined with ST by oversampling the first split; no sampling method is applied to the second split. For each configuration, five independent trial runs were trained with different initial weights. This setting ensures the effect of weight initialization to be ruled out in evaluating model performance. Among the five trials, the model with the highest validation score was used for evaluation. All the experimental settings including epochs, learning rate, and model architecture are fixed for each corresponding task.

4.3 EXPERIMENT RESULTS

Table 2: Experimental results on IMDB

Ratios	$\rho = 10$			$\rho = 20$			$\rho = 50$		
Metrics	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline	0.7291	**0.8751**	0.6248	0.3216	0.6677	0.2118	0.0833	0.69	0.0443
ROS	0.7548	0.702	0.8162	0.5738	0.5839	0.564	0.5837	0.5618	0.6061
RUS	0.7921	0.7305	0.865	0.1837	0.6758	0.1063	0.0256	0.6065	0.0131
ST	0.7956	0.7499	0.8472	0.5844	0.5984	0.571	0.5601	0.5266	0.5981
ST + EWC	0.8002	0.7181	**0.9035**	**0.7543**	**0.7368**	0.7727	0.6457	**0.7188**	0.586
ST + EWC + ROS	**0.8143**	0.7984	0.8307	0.7259	0.6728	**0.788**	**0.6471**	0.6156	**0.6819**

IMDB Results. Table 2 shows the experimental results on the IMDB dataset (Maas et al., 2011). We observe training the model with ST outperforms, if not on par with, traditional methods. Baseline is a setting where no deep learning techniques are employed therefore the learner is trained from the intrinsic imbalanced distribution. Results show a considerable increase in recall when ST is applied. This is because the model is able to predict more positive(minority) instances correctly since focus was put on the under-represented class. It is natural to expect a significant decrease in precision since the focus has been shifted away from the majority class. However, the drop is minimal when EWC is applied. This is because EWC helps the model to remember valuable information obtained during the training of the first split as the model is trained with the balanced second split. It was further observed that applying EWC, ROS, and ST together significantly outperforms other methods. Lastly, as the severity of data imbalance increases, the performance gap between applying ST and not applying grows substantially. In the case of ρ=10 and ρ=50, applying ROS together with ST and EWC shows the best performance in terms of F1-Score.

Table 3: Experimental results on image data

Ratios	(CIFAR-10) $\rho = 5$ $\mu = 0.6$			(CIFAR-10) $\rho = 10$ $\mu = 0.8$			(MNIST) $\rho = 10$ $\mu = 0.8$		
Metrics	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Baseline	0.6695	0.6187	**0.8794**	0.6333	0.5903	0.8422	0.8931	0.8897	0.8966
ROS	0.8187	0.7956	0.8540	0.8194	0.7960	0.8525	0.8984	0.8978	0.8990
RUS	0.8263	0.8028	0.8567	0.8294	0.8073	0.8556	0.8263	**0.8981**	0.8990
ST	0.7916	0.7633	0.8329	0.7901	0.7495	0.8428	0.8962	0.8941	0.8984
ST + EWC	0.8033	0.7733	0.8403	0.8260	0.8045	0.8511	0.8968	0.8950	0.8987
ST + EWC + ROS	**0.8298**	**0.8122**	0.8532	**0.8311**	**0.8099**	**0.8566**	**0.8986**	0.8980	**0.8992**

CIFAR-10 and MNIST Results. Table 3 shows the results of experiments performed on the two image datasets. While MNIST (Lecun et al., 1998) is a simple handwritten digits dataset without color information, CIFAR-10 (Krizhevsky, 2012) is a relatively complicated image classification task with color information. The experimental results manifest larger gap exists when using ST on CIFAR-10 rather than on MNIST. This accounts to the relative simplicity of MNIST dataset which alleviates the shortcoming of skewed data during representation learning. This shows that ST greatly benefits the learner in a difficult setting where the the task difficulty and imbalance ratio is severe. The results show that a variant of our proposed method, ST with EWC and ROS, scores the best among other approaches in general.

5 CONCLUSION

Learning from imbalanced class inevitably brings bias toward frequently observed classes. Data-level manipulation tries to under-sample the majority classes or over-sample the minority classes. But these methods have a tendency to discard valuable information from observations of majority classes or overfit to sparse representation of minority classes, especially as the imbalance level gets higher. If the learning of a classification model is limited to maximizing the total accuracy over the entire data, models pay more attention to majority classes while neglecting the rest.

We propose Sequential Targeting, which effectively circumvents these issue by simply decomposing the data into k splits and sequentially training a learner in the decreasing order of KL divergence with the target distribution, which in the case of data imbalance problem is the discrete uniform distribution. Our architecture proves to be compatible with previous methods and outperforms existing methods when validated on imbalanced text and image classification tasks. Our model shows superiority in performance because of simultaneous increase in both precision and recall thereby improving the overall F1-score. We believe that our work makes a meaningful step towards the application of incremental learning on the data imbalance problem.

6 DISCUSSION

Variations of η ratio has been tested. However, it proves to be domain-dependent and most variations still outperformed previous methods. A fixed ratio of 1:1 was used throughout the experiment in this paper. However, in order to fully utilize ST, variations of different η ratio should be tested for optimal performance.

In the case of ROS and RUS, ensemble methods could be utilized since data is randomly sampled. Each sample instance can be considered to train weak learners that can use max-voting schemes to create a single strong learner. Likewise, variations of η ratio can be considered different samples train multiple weak learners. In further research, ensemble methods of ROS, RUS, and ST will be tested as well.

Lastly, since ST is compatible with algorithm-level methods, successful methods such as focal loss (Lin et al., 2017) and cost-sensitive deep neural network (Khan et al., 2015) are expected to increase overall performance if implemented together.

ACKNOWLEDGMENTS

This work was funded by NAVER Corp.

REFERENCES

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. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2018.07.011. URL http://www.sciencedirect.com/science/article/pii/S0893608018302107.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority oversampling technique. *JAIR Journal of Artificial Intelligence Research*, 16, 2002. doi: https://doi.org/10.1613/jair.953.

Nitesh Chawla, Aleksandar Lazarevic, Lawrence Hall, and Kevin Bowyer. Smoteboost: Improving prediction of the minority class in boosting. volume 2838, pp. 107–119, 01 2003. doi: 10.1007/978-3-540-39804-2_12.

D. A. Cieslak, N. V. Chawla, and A. Striegel. Combating imbalance in network intrusion datasets. In *2006 IEEE International Conference on Granular Computing*, pp. 732–737, 2006.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. 2017.

Robert French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3: 128–135, 05 1999. doi: 10.1016/S1364-6613(99)01294-2.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. 2013.

Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. pp. 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

Khoshgoftaar T.M. Johnson, J.M. Survey on deep learning with class imbalance. *J Big Data*, 6, 2019. doi: https://doi.org/10.1186/s40537-019-0192-5.

Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Cost sensitive learning of deep feature representations from imbalanced data, 2015.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *ICML*, 18:12–13, 2016.

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell*, 5:221–232, 2016. doi: https://doi.org/10.1007/s13748-016-0094-0.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct 2017.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

David Masko and Paulina Hensman. The impact of imbalanced training data for convolutional neural networks. 2015.

Kamthorn Puntumapon, Thanawin Ralthamamom, and Kitsana Waiyamai. Cluster-based minority over-sampling for imbalanced datasets. *IEICE Transactions on Information and Systems*, E99.D (12):3101–3109, 2016. doi: 10.1587/transinf.2016EDP7130.

R. Bharat Rao, Sriram Krishnan, and Radu Stefan Niculescu. Data mining for improved cardiac care. *SIGKDD Explor. Newsl.*, 8(1):3–10, June 2006. doi: 10.1145/1147234.1147236.

Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks, 2016.

Victor Sheng and Charles Ling. Thresholding for making classifiers cost sensitive. volume 1, 01 2006.

Li J. Cao L. et al. Wei, W. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16:449–475, 2013. doi: https://doi.org/10.1007/s11280-012-0178-0.