1 A Quadratic Mean Based Supervised Learning Model for Managing Data Skewness

This chapter reviews a novel approach for addressing data skewness in supervised learning models, presented in the paper "A Quadratic Mean Based Supervised Learning Model for Managing Data Skewness." The authors propose a framework called QMLearn, which introduces a new way of calculating empirical risk using the quadratic mean, aimed at improving model robustness on imbalanced datasets.

1.1 Introduction to the Problem of Data Skewness

Data skewness, or imbalance, is a prevalent issue in supervised learning where the distribution of the dependent variable is uneven. This imbalance often causes traditional models to become biased towards the majority class, resulting in poor performance on the minority class. This paper identifies the limitations of traditional empirical risk minimization methods and introduces QMLearn to better handle skewed data.

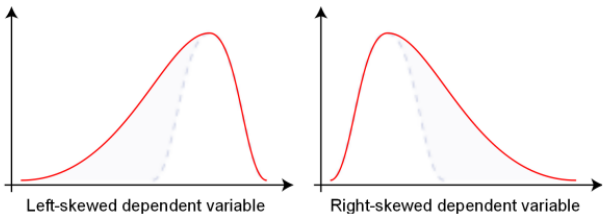


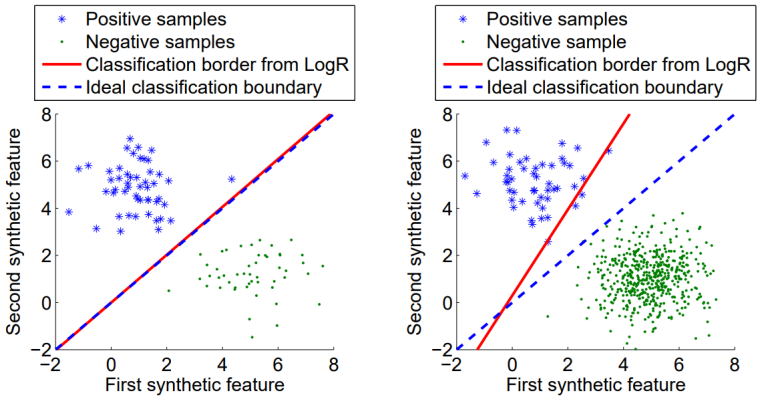
Figure 1: Data skewness.

1.2 Limitations of Traditional Learning Models

Traditional learning models, such as logistic regression and SVMs, typically minimize an empirical risk function defined as the arithmetic mean of the loss across all training examples:

$$R_{\text{emp}}(w) = \frac{1}{n} \sum_{i=1}^{n} l(x_i, y_i, w) \tag{1}$$

where n is the total number of training instances, x_i represents the feature vector, y_i the true label, and w the model parameters. This method often results in models that are biased towards the majority class, as illustrated in Figure 2.



(a) Class ratio = 1:1 (b) Class ratio = 1:10

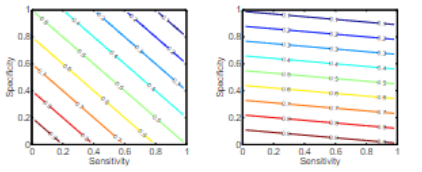
Figure 2: Classification boundaries using traditional logistic regression on balanced (left) and imbalanced (right) datasets.

1.3 Introduction to the QMLearn Framework

The QMLearn framework modifies the traditional empirical risk by using the quadratic mean rather than the arithmetic mean. This redefinition is designed to address the imbalance by equalizing the influence of both classes on the model's training process. The quadratic mean-based empirical risk function is defined as:

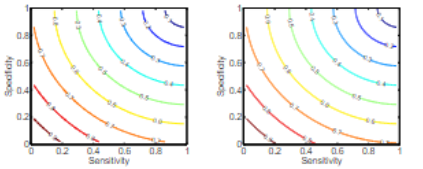
$$R^{Q}_{\text{emp}}(w) = \sqrt{\frac{\left(\frac{\sum_{i=1}^{n_1} l(x_i, y_i, w)}{n_1}\right)^2 + \left(\frac{\sum_{i=n_1+1}^{n} l(x_i, y_i, w)}{n_2}\right)^2}{2}} \tag{2}$$

where n_1 and n_2 are the number of instances in each class. This approach balances the error contributions from both classes, making the model more robust against skewed distributions.



(a) Original empirical risk on balanced data (b) Original empirical risk on imbalanced data (Pos:Neg = 1:10)

Figure 3: empirical risk on balanced (left) and imbalanced (right) data.



(a) **QMLearn-based empirical risk on balanced data** (b) **QMLearn-based empirical risk on imbalanced data (Pos:Neg = 1:10)**

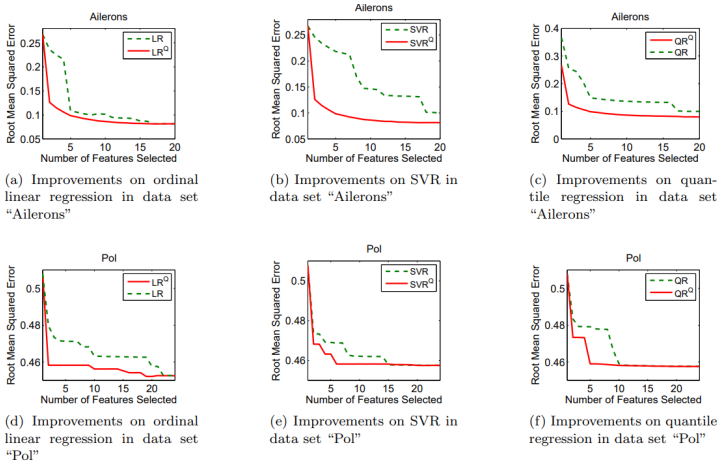
Figure 4: QMLearn-based empirical risk on balanced (left) and imbalanced (right) data.

1.4 Convex Optimization for Efficient Learning

The QMLearn method is formulated as a convex optimization problem, maintaining the convexity of the empirical risk function through the use of the quadratic mean. This allows for efficient computation and ensures that the solution is optimal, even for large-scale datasets.

1.5 Experimental Validation

Extensive experiments were conducted to validate the effectiveness of QMLearn compared to traditional models like logistic regression, SVMs, and quantile regression. The results, depicted in Figure 5, show that QMLearn consistently outperforms traditional methods, particularly on imbalanced datasets.



(a) Improvements on ordinal linear regression in data set "Ailerons"

(b) Improvements on SVR in data set "Ailerons"

(c) Improvements on quantile regression in data set "Ailerons"

(d) Improvements on ordinal linear regression in data set "Pol"

(e) Improvements on SVR in data set "Pol"

(f) Improvements on quantile regression in data set "Pol"

Figure 5: Performance improvements of QMLearn-based models in regression tasks on datasets 'Ailerons' and 'Pol'.

1.6 Using the QMLearn Package

To implement the QMLearn framework in your own work, you can install the QMLearn package by following the instructions available at qmlearn.rutgers.edu/source/install.html. For more details on installation and usage, please refer to the official QMLearn documentation.

1.7 Conclusion

The QMLearn framework offers a robust solution for managing data skewness in supervised learning models. By redefining the empirical risk function using the quadratic mean, QMLearn ensures balanced error consideration across classes, improving model performance on imbalanced datasets. Future research could explore the application of QMLearn to other types of models and further investigate its theoretical underpinnings.

2 Machine Learning for Resource Estimation in Highly Skewed Gold Deposits

This chapter explores the methodologies proposed in the paper "A Novel Approach for Resource Estimation of Highly Skewed Gold Using Machine Learning Algorithms." The authors introduce a machine learning-based framework to address the challenges posed by highly skewed distributions of gold grades in vein deposits. Traditional geostatistical methods, such as ordinary kriging and indicator kriging, often struggle to model these distributions accurately due to their reliance on assumptions of normality and spatial continuity. The proposed machine learning approaches offer a more flexible and robust alternative for estimating resources in such complex geological settings.

2.1 Understanding Skewed Vein Deposits

Vein deposits often exhibit highly skewed distributions of gold grades, primarily due to the nugget effect, which causes significant variability in gold concentration. In these deposits, the majority of samples have low gold grades, while a few samples exhibit very high grades. This creates a long-tailed, or right-skewed, distribution, which complicates the accurate estimation of gold resources. The challenge lies in developing models that can accurately predict both the frequent low-grade occurrences and the rare high-grade values without bias.

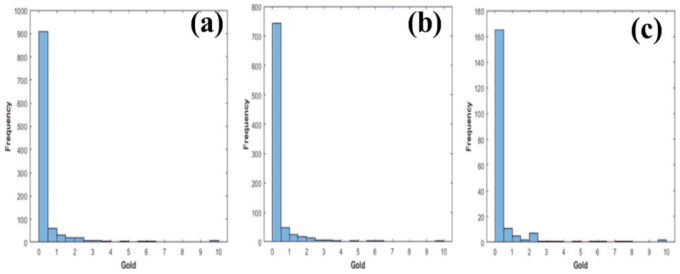


Figure 6: Histogram of gold values for (a) entire dataset, (b) training dataset, and (c) testing dataset. The histograms illustrate the right-skewed nature of the gold distribution, with a high frequency of low-grade values and a few high-grade values.

2.2 Data Normalization Techniques

To effectively handle the skewed distribution of gold grades, the authors applied two normalization techniques: logarithmic normalization and z-score normalization. These techniques are crucial for transforming the data into a form that is more suitable for machine learning models, particularly when dealing with highly skewed data.

2.2.1 Logarithmic Normalization

Logarithmic normalization is a transformation that reduces the impact of outliers by compressing the range of values. This technique is particularly effective for right-skewed data, as it mitigates the effect of extreme high-grade values, thereby creating a more balanced distribution. By applying a logarithmic transformation, the differences between high-grade and low-grade values are reduced, making the data more symmetric and better suited for regression modeling. This transformation is defined as:

$$x' = \log(x + 1)$$

where x is the original data point, and x' is the transformed data. The addition of 1 ensures that zero values are handled appropriately, as the logarithm of zero is undefined.

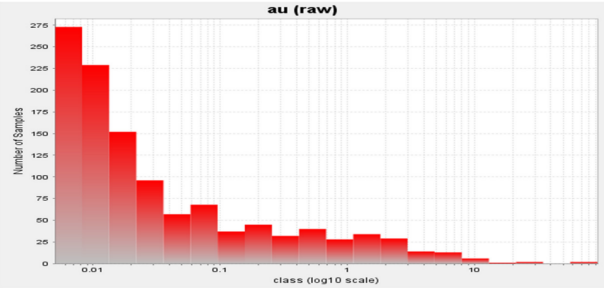


Figure 7: Logarithmic histogram of gold distribution, showing how logarithmic normalization reduces skewness by transforming the original right-skewed data into a more balanced distribution.

2.2.2 Z-Score Normalization

Z-score normalization standardizes the data by subtracting the mean and dividing by the standard deviation. This technique centers the data around zero with a unit variance, stabilizing the variance across different scales of data. Z-score normalization is effective when the data distribution is approximately normal. However, when applied to skewed data, it can still be useful by scaling the data in a way that allows machine learning models to better capture the underlying patterns. The z-score normalization formula is:

$$z = \frac{x - \mu}{\sigma}$$

where x is the data point, μ is the mean of the dataset, and σ is the standard deviation.

2.3 Data Segmentation Using Marine Predators Algorithm (MPA)

To ensure that both high-grade and low-grade gold samples are adequately represented in the training and testing datasets, the authors employed the Marine Predators Algorithm (MPA) for data segmentation. MPA is an optimization algorithm inspired by the foraging strategies of marine predators. It is used to create balanced datasets by effectively capturing the distribution of gold grades across different segments. This segmentation approach prevents the model from being biased towards either end of the spectrum, enhancing its accuracy and generalization capabilities.

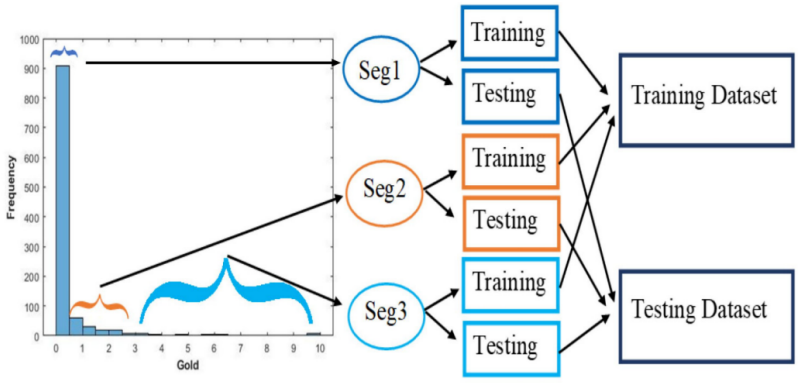


Figure 8: Data segmentation of gold values for training and testing based on the histogram plot. The Marine Predators Algorithm (MPA) was used to create balanced training and testing datasets, effectively capturing the distribution of gold grades across different segments.

2.4 Machine Learning Models for Estimation

The study explored several machine learning algorithms, including Gaussian Process Regression (GPR), Support Vector Regression (SVR), Decision Tree Ensemble (DTE), Fully Connected Neural Network (FCNN), and K-Nearest Neighbors (K-NN). Each model has unique strengths and weaknesses in handling skewed data distributions, making them suitable for different aspects of the resource estimation task.

2.4.1 Gaussian Process Regression (GPR)

GPR is a non-parametric, probabilistic model that provides a flexible framework for regression tasks. It models the distribution of the target variable and its uncertainty, making it well-suited for handling the variability in gold grades. GPR can capture complex, non-linear relationships in the data, which is particularly important for accurately predicting rare, high-grade values.

2.4.2 Support Vector Regression (SVR)

SVR uses kernel functions to handle non-linear relationships in the data by maximizing the margin of error while minimizing model complexity. This approach is effective in managing skewness and outliers, as it focuses on a subset of critical points (support vectors) that define the model, rather than all the data points.

2.4.3 Decision Tree Ensemble (DTE)

Ensemble methods like Random Forests or Gradient Boosting Machines aggregate multiple decision trees to improve prediction accuracy and robustness. These models can handle varying distributions within the data by creating diverse decision paths, making them suitable for estimating resources in skewed deposits.

2.4.4 Fully Connected Neural Network (FCNN)

Neural networks can learn complex patterns and relationships in the data through multiple layers of interconnected neurons. FCNNs are particularly effective for capturing non-linear dependencies and can adapt to skewed distributions by learning directly from the data without assuming any prior distribution.

2.4.5 K-Nearest Neighbors (K-NN)

K-NN is a non-parametric method that predicts output values based on the average output of the k-nearest training samples. It is capable of handling skewed data by considering local structures rather than global distribution assumptions, making it a useful tool for resource estimation in complex geological settings.

2.5 Performance Evaluation

The authors evaluated the performance of these machine learning models using both normalized datasets (z-score and logarithmic) to handle the skewness effectively. The results demonstrated that machine learning models could predict gold grades accurately, even in the presence of highly skewed data.

2.5.1 Performance after Logarithmic Normalization

After applying logarithmic normalization, the machine learning models showed high accuracy in capturing the variability of gold grades, effectively handling the skewed data. This approach allowed the models to generalize well across different gold grades, as shown in the predicted versus actual plots.

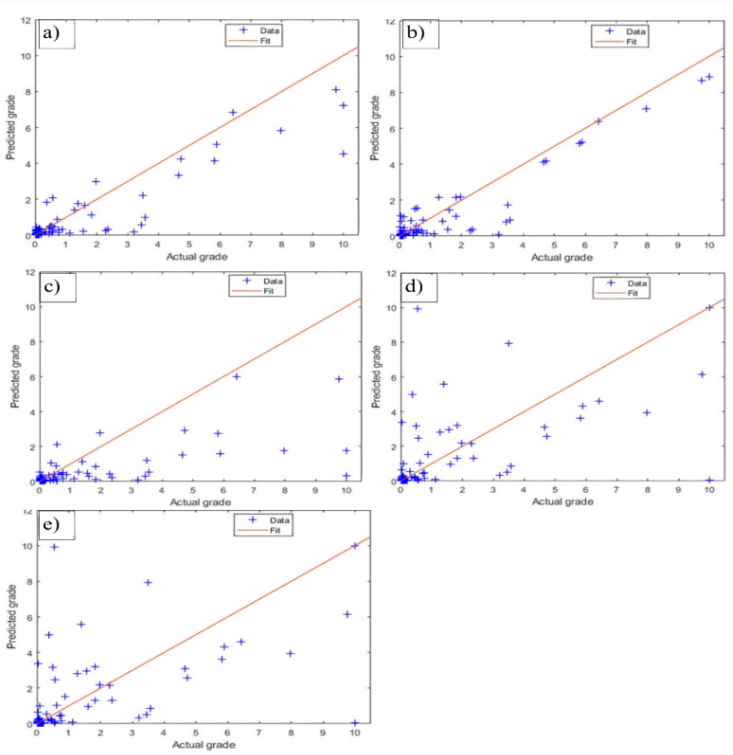


Figure 9: Predicted vs. actual gold grades after logarithmic normalization for (a) Gaussian Process Regression (GPR), (b) Support Vector Regression (SVR), (c) Decision Tree Ensemble (DTE). The models show good agreement with the actual values, indicating successful handling of the skewed data.

2.5.2 Performance after Z-Score Normalization

Similarly, z-score normalization also provided effective results, particularly for models that assume normal distribution of data. The prediction accuracy remained high across different gold grades, showcasing the robustness of machine learning models in resource estimation tasks.

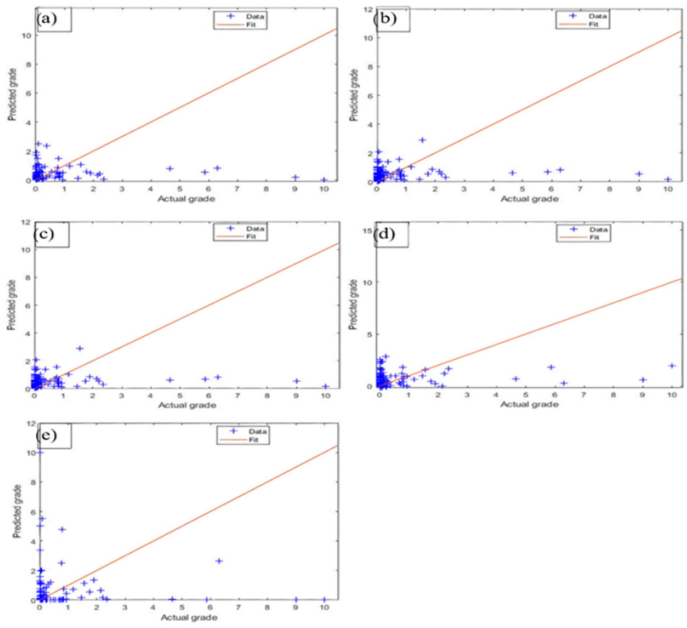


Figure 10: Predicted vs. actual gold grades after z-score normalization for (a) Gaussian Process Regression (GPR), (b) Support Vector Regression (SVR), (c) Decision Tree Ensemble (DTE). The plots demonstrate the models' capability to predict gold grades accurately, maintaining consistency across different normalization techniques.

2.6 Conclusion

The paper presents a robust methodology for estimating gold resources in highly skewed vein deposits using machine learning algorithms. By employing normalization techniques and advanced data segmentation strategies, the study highlights the potential of machine learning to outperform traditional geostatistical methods in complex geological settings. The findings suggest that machine learning models, particularly when optimized with appropriate preprocessing and data handling techniques, can provide accurate and reliable resource estimates even in the presence of significant data skewness. Future research may focus on refining these models further and exploring their application to other types of mineral deposits.

3 Distributional Robustness Loss for Long-Tail Learning

This chapter summarizes the methodologies proposed in the paper "Distributional Robustness Loss for Long-Tail Learning" by Dvir Samuel and Gal Chechik. The paper addresses the challenge of learning from unbalanced datasets with long-tailed distributions, where a few classes have a large number of samples (head classes), while many classes have very few samples (tail classes). The authors introduce a novel loss function based on Distributionally Robust Optimization (DRO) to improve the learning of representations for both head and tail classes in deep learning models.

3.1 Challenges of Long-Tail Learning

In real-world datasets, long-tailed distributions are common, where the frequency of classes follows a steep drop-off from the head to the tail. This imbalance causes deep models to be biased towards head classes, leading to poor recognition performance for tail classes. Traditional approaches to handling unbalanced data, such as data resampling, loss reweighting, and classifier adjustments, primarily focus on balancing the classifier's output. However, these methods do not adequately address the underlying representation learned by the model, which remains biased towards the head classes.

3.2 Distributionally Robust Optimization (DRO)

To mitigate the biases in representation learning, the authors propose using Distributionally Robust Optimization (DRO). DRO aims to minimize the worst-case loss within an uncertainty set of possible distributions, thereby enhancing the model's robustness to shifts in the data distribution. The DRO framework is particularly suitable for long-tail learning, as it allows the model to better generalize to tail classes, which are underrepresented in the training data.

The DRO problem is formulated as follows:

$$\text{DRO} : \min_{f} \sup_{Q \in \mathcal{U}} E_{(x,y)\sim Q}[l(f(x), y)] \tag{3}$$

Here, f represents the model, $\mathcal{U}$ is an uncertainty set around the empirical training distribution $\hat{P}$, and $l(f(x), y)$ is the loss function. The goal is to minimize the worst-case expected loss over all distributions Q within the uncertainty set $\mathcal{U}$.

3.3 Distributional Robustness Loss (DRO-LT)

The authors introduce a novel loss function, called DRO-LT Loss, designed to improve representation learning under long-tailed data distributions. This loss function extends standard contrastive losses, which pull samples closer to the

centroid of their own class and push away samples from other classes. DRO-LT Loss, however, accounts for the uncertainty in estimating class centroids, particularly for tail classes with fewer samples.

The DRO-LT loss function is defined as:

$$L_{\text{Robust}} = -\sum_{c \in C} w(c) \sum_{z \in S_c} \log \frac{e^{-d(\hat{\mu}_c, z) - 2\epsilon_c}}{\sum_{z' \in Z} e^{-d(\hat{\mu}_c, z') - 2\epsilon_c \delta(z', c)}} \tag{4}$$

where:

- $\hat{\mu}_c$ is the empirical centroid of class c. - $d(\hat{\mu}_c, z)$ is the distance between the feature representation z and the centroid $\hat{\mu}_c$. - ϵ_c is the robustness margin for class c, which accounts for the uncertainty in the centroid estimation. - $\delta(z', c) = 1$ if z' is of class c, and 0 otherwise. - $w(c)$ are class weights that ensure balanced contributions from all classes.

3.4 Optimization Strategy

To compute the DRO-LT loss, an initial feature representation of the data is required for estimating class centroids. The training process involves two stages:

1. **Initial Training**: The model is first trained using standard cross-entropy loss to learn initial feature representations and centroids. 2. **Robust Training**: The DRO-LT loss is then introduced by combining it with the standard cross-entropy loss, allowing the model to learn robust representations that generalize well to tail classes.

The combined loss function used for training is:

$$L = \lambda L_{\text{CE}} + (1 - \lambda) L_{\text{Robust}} \tag{5}$$

where L_{CE} is the standard cross-entropy loss, and λ is a trade-off parameter that balances the two loss components.

3.5 Empirical Results

The authors evaluated their approach on several long-tailed visual recognition benchmarks, including CIFAR100-LT, ImageNet-LT, and iNaturalist. The results demonstrate that the DRO-LT loss consistently outperforms state-of-the-art methods in long-tail learning. It improves recognition accuracy for tail classes while maintaining high accuracy for head classes, showcasing the effectiveness of the proposed method in learning balanced representations.

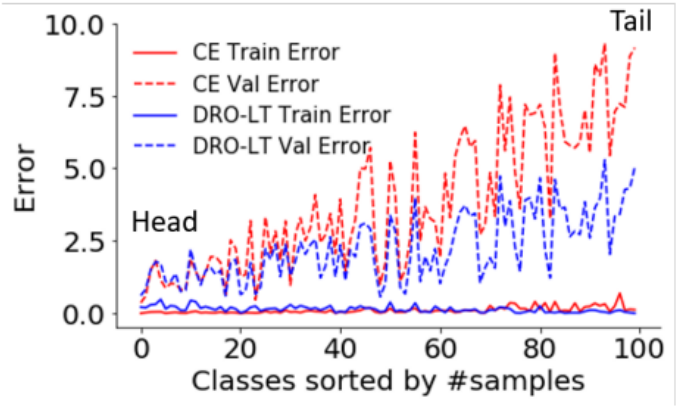


Figure 11: Performance comparison of DRO-LT with state-of-the-art methods on long-tailed benchmarks. The DRO-LT method achieves superior accuracy for both head and tail classes.

3.6 Conclusion

The paper introduces a novel approach for long-tail learning by focusing on distributionally robust optimization. The DRO-LT loss effectively addresses the challenges of unbalanced data by improving the learned representations for both head and tail classes. This robust representation learning method sets new state-of-the-art results on multiple benchmarks, indicating its potential for broader applications in unbalanced data scenarios.