Difficult Samples Re-embedding via Mutual Information Constrained Semantically Oversampling

Jiachen Tian, Shizhan Chen, Xiaowang Zhang*, Zhiyong Feng, Deyi Xiong, Shaojuan Wu and **Chunliu Dou**
College of Intelligence and Computing, Tianjin University, Tianjin, China
{jiachen6677, shizhan, xiaowangzhang, zyfeng, dyxiong, shaojuanwu, 2019229041}@tju.edu.cn

Abstract

Difficult samples of the minority class in imbalanced text classification are usually hard to be classified as they are embedded into an overlapping semantic region with the majority class. In this paper, we propose a Mutual Information constrained Semantically Oversampling framework (MISO) that can generate anchor instances to help the backbone network determine the re-embedding position of a non-overlapping representation for each difficult sample. MISO consists of (1) a semantic fusion module that learns entangled semantics among difficult and majority samples with an adaptive multi-head attention mechanism, (2) a mutual information loss that forces our model to learn new representations of entangled semantics in the non-overlapping region of the minority class, and (3) a coupled adversarial encoder-decoder that fine-tunes disentangled semantic representations to remain their correlations with the minority class, and then using these disentangled semantic representations to generate anchor instances for each difficult sample. Experiments on a variety of imbalanced text classification tasks demonstrate that anchor instances help classifiers achieve significant improvements over strong baselines.

1 Introduction

Data imbalance is a long-standing challenge in the text classification tasks such as sentiment analysis (Wu et al., 2018), intent detection (Quan et al., 2020) and spam detection (Liu et al., 2017), where the distribution of training data over classes is skewed. For example, the number of minority samples accounts for only 28% of training instances in SMS Spam dataset (Peng et al., 2019) and 14% in Opin-Rank dataset (Ganesan and Zhai, 2012). Data imbalanced issue is more severe in Toutiao dataset (Ouyang et al., 2020) with a minority-majority ratio of 1:122 (hereafter imbalance ratio).

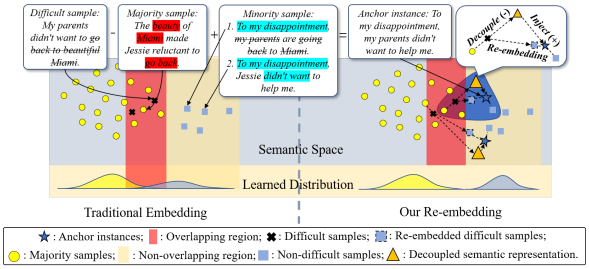

Figure 1: Visualization of the data imbalance problem.

As the class distribution tends to be extremely imbalanced, texts from the minority class(es) may be easily categorized into the majority class(es) (He and Garcia, 2009; Fernández et al., 2018; Gao et al., 2020; Yang et al., 2020).

Datasets	Perc. (%) of DS | Non-DS	F1-score (%) of DS | Non-DS
Opin-Rank	15.3 | 84.7	56.2 | 99.3
SMS Spam	42.1 | 57.9	64.3 | 94.6
Toutiao	36.6 | 63.4	59.0 | 95.6
Yelp.P (1%)	28.2 | 71.8	32.6 | 94.3
Yelp.P (5%)	22.4 | 77.6	46.4 | 93.8
IMDB (1%)	49.3 | 50.7	18.8 | 83.6
IMDB (5%)	30.1 | 69.9	26.9 | 87.3
AG_News (1%)	48.5 | 51.5	17.5 | 75.9
AG_News (5%)	33.2 | 66.8	27.7 | 83.8

Table 1: Statistics with respect to imbalanced datasets, and the classification performance of XLNet about difficult and non-difficult samples. DS: difficult samples. Non-DS: non-difficult samples. Perc.: percentages.

Recent studies have shown that some minority samples, called difficult samples as they locate in the overlapping semantic region, are more important for imbalanced text classification than those far from the overlapping semantic region (Girshick et al., 2014; Robinson et al., 2020). As illustrated in Figure 1, the difficult samples have similar (entangled) embeddings with some majority samples in this overlapping semantic region as they are similar to these majority samples about surface forms (e.g., n-gram or syntax). For example, in Yelp.P dataset, a review "*my parents didn't want to go back to beautiful Miami*" is a difficult sample of the minority class. However, many words of this sample have

*Corresponding author: xiaowangzhang@tju.edu.cn

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3148–3161
November 7–11, 2021.

also occurred in a positive review of the majority class "*the beauty of Miami made Jessie reluctant to go back*". Table 1 shows the percentages of difficult samples in several imbalanced datasets and the impact of difficult samples on classification performance of the strongest baseline (XLNet). Clearly, Classification errors mainly come from the misclassification of difficult samples. The most serious situation appears in AG_News (1%) dataset, where difficult samples account for 48.5% of minority samples, and XLNet only obtained 17.5% F1-score for these difficult samples.

The latest research on imbalanced learning separated the learning procedure into representation learning (i.e., the backbone network) and classification (i.e., the classifier), and achieved the state-of-the-art performance by freezing the backbone network and fine-tuning the classifier weights to obtain balanced decision boundaries (Kang et al., 2020). However, they ignore that entangled semantic representations of difficult samples make the decision boundaries hard to be clearly determined.

To this end, we propose to generate anchor instances, which have similar surface forms with difficult samples but be embedded in the non-overlapping region of the minority class, to help the backbone network learn disentangled semantic representations for difficult samples. See Figure 1, consider the aforementioned difficult sample, two anchor-instance-generation steps are taken. First, entangled semantics of "*beauty*", "*Miami*" and "*go back*" in the difficult sample, are decoupled from the anchor instance. Second, semantics of "*to my disappointment*" and "*help me*" in some non-difficult samples of the minority class are injected into the anchor instance.

In order to make this generation framework feasible, we should answer the following three questions: (1) Given a difficult sample paired with a majority sample, how can we capture their entangled semantics? (2) How can we decouple and inject semantics from and into an anchor instance? (3) Merging anchor instances into the original data may change the data distribution and hence have a negative impact on non-difficult sample classification. How can we avoid this?

To address these problems, we propose a **M**utual **I**nformation constrained **S**emantically **O**versampling (MISO) approach with three essential components: a semantic fusion module (SFM), a mutual information (MI) loss, and a coupled adversarial generator (CAG) based on encoder-decoder networks. SFM is leveraged to adaptively find entangled semantics among difficult samples and majority samples in the overlapping region. Formally, we assume the majority and minority classes as two random variables A and B, their entangled semantics can be modeled as the mutual information between the two classes: $\mathcal{I}(A;B) = \sum_{b\in B}\sum_{a\in A} p(a,b)\log\frac{p(a,b)}{p(a)p(b)}$. We introduce MI loss for parallel decoupling and injecting, which is symmetric and smooth. Semantic representations outputted from SFM constrained by MI loss are fed into CAG for generating anchor instances, with an adversarial strategy to ensure that the original data distribution is not destroyed.

In addition to the proposed MISO framework for imbalanced text classification, other contributions of our work can be summarized as follows. First, the boundary of the minority class learned by MI loss is theoretically proved as a (near-)optimal boundary. Second, we further theoretically show that the new distribution after adding anchor instances is consistent with the original distribution of the minority class (see proof.1 and proof.2 in Appendix). Third, experiment results demonstrate that text classifiers, trained on rebalanced datasets with anchor instances generated by MISO, outperform state-of-the-art methods by an average of 2.7% in nine datasets. The average success rate of moving difficult samples into non-overlapping region is 13.7%, which validates the effectiveness and robustness of MISO in handling difficult samples.

2 Related Work

Imbalanced Learning in NLP The re-sampling approach to this issue restores the balance of the class distribution by either undersampling the majority class or oversampling the minority class (Han et al., 2005; Chawla et al., 2002; Cao et al., 2019). Cost-sensitive methods estimate the cost of samples with a cost matrix and train the classifier with different penalties (Gomez et al., 2000; McBride et al., 2019). Additionally, text style transfer with generative adversarial networks (GANs) has been used for oversampling, too (Fu et al., 2018; Guo et al., 2018; Nie et al., 2019). One advantage of these methods is that generated texts still follow the original data distribution. Kang et al. (2020) propose a long-tailed learning approach (τ-norm and cRT) to separate representation learning and classifier training. Chen et al. (2020) introduce MixText

with TMix, a data augmentation method similar to Mixup used in computer vision, to interpolate new points in their corresponding hidden space.

Difficult Sample Modeling in NLP Lin et al. (2017) propose a soft sampling method that dynamically adjusts the weights of difficult samples by redefining the loss function. Dice loss that optimizes the Sørensen–Dice coefficient to immune the imbalance issue has also been proposed (Li et al., 2020). Glazkova (2020) introduces ADASYN to assign a weight for each minority instance.

Difficult Sample Modeling in CV Difficult sample learning is one of fundamental issues in object dectection (Oksuz et al., 2019). Inspired by the view that difficult samples are usually with a high loss, several studies adopt a bootstrapping to mine difficult samples (Felzenszwalb et al., 2009; Ren et al., 2015). GANs are also used to generate difficult samples (Wang et al., 2017). Pang et al. (2019) propose a method based on Intersection-over-Unions to sample negative examples.

Our Proposal Significantly different from previous methods, our proposed MISO explores mutual information to decouple the overlapping between the majority and minority classes, which theoretically guarantees the consistency of class distribution after oversampling.

3 Problem Statement

Let $\mathbf{X}^+ := \{x_1^+, ..., x_{n^+}^+\} \in \mathbb{R}^{n^+ \times l}$ be a training set of positive samples with the minority class distribution $\mathbb{N}^+$, $\mathbf{X}^- := \{x_1^-, ..., x_{n^-}^-\} \in \mathbb{R}^{n^- \times l}$ be a training set of negative samples with the majority class distribution $\mathbb{N}^-$, where x_i is the i-th sentence consisting of up to l tokens, n^- and n^+ are the number of instances in the majority and minority classes, respectively. Data imbalance can be roughly divided into the slight imbalance (e.g., $\frac{n^+}{n^-} = \frac{4}{6}$) and the severe imbalance (e.g., $\frac{n^+}{n^-} = \frac{1}{100}$ or less) (He and Garcia, 2009; Brownlee, 2019).

MISO learns a joint distribution $\mathbb{Z}$ for the majority and minority classes in the same semantic space. From this distribution, we sample $\mathbf{Z} := \{z_1, ..., z_m\} \in \mathbb{R}^{m \times d}$, which consists of $m \in [0, n^+ \times n^-]$ d-dimensional vectors.

The goal of MISO is to make $\mathbb{Z}$ close to $\mathbb{N}^+$ but far from $\mathbb{N}^-$. In doing so, we generate a set of anchor instances $\mathbf{Y}^+ := \{y_1^+, ..., y_t^+\} \in \mathbb{R}^{t \times l}$ with $\mathbf{Z}$ as their disentangled representations for difficult samples in $\mathbf{X}^+$, where t is the number of anchor instances. We further define a marginal distribution over $\mathbf{Y}^+$ as $\mathbb{U}_{\psi_+,\sigma,\omega}$, where ψ_+, σ, ω are the parameters of a continuous and differentiable parametric function E_{ψ_+} (i.e., the minority encoder-decoder), SFM, and MI loss.

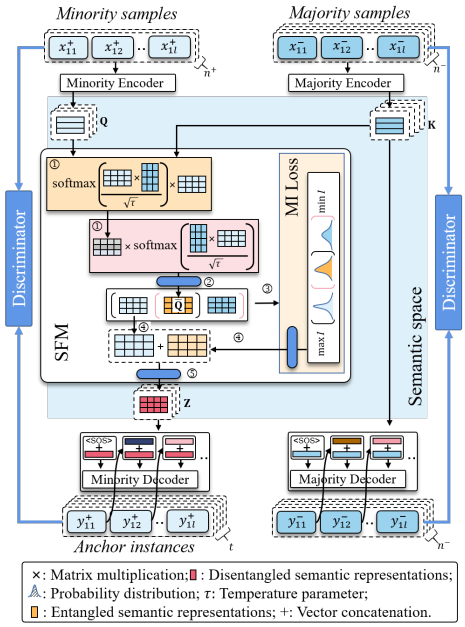

Figure 2: Architecture of the proposed mutual information constrained semantically oversampling framework.

4 MISO

We introduce the overall architecture and then elaborate on each component of MISO in this section.

4.1 The Overall Architecture

As shown in Figure 2, MISO is built upon a coupled adversarial encoder-decoder framework that consists of two encoders together with two decoders (i.e., a latent variable-guided decoder and a standard one) and two discriminators. The two encoders are used to encode instances from the minority class (left encoder) and instances from the majority class (right encoder). To project instances from the two classes into the same semantic space, the two encoders share their parameters. MISO is equipped with two additional components: SFM and MI loss.

SFM captures the entangled semantics of difficult samples by extracting semantics of the minority class that is similar to those of the majority class (Step. ①). Learned entangled semantic representations are fused into a feedforward layer (Step. ②) and then fed into two Mutual Information Neural Estimators (MINEs) (Belghazi et al.,

2018) (Step. ③). MI loss uses these MINEs to decouple entangled semantic representations from the majority class (Step. ④) and then feed these disentangled semantic representations into another feedforward layer (Step. ⑤). Specifically, MI loss minimizes the mutual information between entangled semantic representations and the minority class at the decoupling step and maximizes the mutual information between disentangled semantic representations and the majority class at the injecting step. In doing so, we move entangled semantic representations from the overlapping region into the non-overlapping region of the minority class, which are disentangled with the majority class.

Disentangled semantic representations are then fed into the minority class decoder (left decoder) to generate anchor instances, which are not hard to classify. The right decoder is used to generate instances of the majority class. Both decoders are monitored by two discriminators that adversarially detects whether the newly generated texts are the same as the original inputs in the surface forms.

4.2 Model Components

SFM In this module, we use a multi-head attention mechanism to learn the entangled semantic part of the input difficult samples.

Each attention head obtain initial semantic representations $\mathbf{Q}$ and $\mathbf{K}$ by calculating $Ee_{\psi_+|}(x^+)$ and $Ee_{\psi_-|}(x^-)$, where $Ee_{\psi_+|}$ and $Ee_{\psi_-|}$ are two encoders with their parameters ψ_+ | and ψ_- |. Once we have $\mathbf{Q}$ and $\mathbf{K}$, we can obtain entangled semantic representations $\bar{\mathbf{Q}}$ as follows:

$$\bar{\mathbf{Q}} = \left[\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\mathsf{T}}{\sqrt{\tau}})\mathbf{Q}\right] \text{softmax}(\frac{\mathbf{Q}^\mathsf{T}\mathbf{K}}{\sqrt{\tau}}),$$

where $\tau = \frac{\#majority\ samples}{\#minority\ samples}$ in the current epoch, so that $\tau \in [1, +\infty)$ is an adaptive temperature parameter to control the scope of entangled semantics that currently needs to be extracted. In other words, in the initial epoch, entangled semantics of difficult samples are difficult to capture, so each difficult sample needs to be compared with (near-)entire majority samples. In the final epoch, the ability of SFM to extract entangled semantics is significantly enhanced, so each difficult sample only needs to be compared with partial majority samples, which have the stronger semantic similarity to this difficult sample. Finally, SFM obtains $b_s \cdot h$ triples $\mathbf{Q}, \bar{\mathbf{Q}}, \mathbf{K}$ after a feed-forward network, where b_s is the size of minibatch, h is the number of attention heads. We denote the distributions of $\mathbf{Q}, \bar{\mathbf{Q}}, \mathbf{K}$ as $\mathbb{Q}, \bar{\mathbb{Q}}$ and $\mathbb{K}$ respectively, and SFM as S_σ with its parameters σ.

MI Loss We propose to use mutual information to calculate semantic similarity because the loss value computed by the mutual information can obtain a (near-)optimal boundary of the minority class. The theoretical proof is shown in the Appendix. We first estimate the mutual information by two MINEs (Belghazi et al., 2018), T_{ω_+} with parameters ω_+, and T_{ω_-} with parameters ω_-. T_{ω_+} is an integrability function to estimate the KL-divergence between the joint distribution $\mathbb{Q}\bar{\mathbb{Q}}$, and the product of the marginals $\mathbb{Q} \otimes \bar{\mathbb{Q}}$. T_{ω_-} is used to estimate the KL-divergence between the joint distribution $\mathbb{K}\bar{\mathbb{Q}}$, and the product of the marginals $\mathbb{K} \otimes \bar{\mathbb{Q}}$. Since KL-divergence can be approximated to a low-bound by its Donsker-Varadhan (DV) representation (Donsker and Varadhan, 1975), both of MINEs are represented as follows:

$$\begin{aligned}
\mathcal{I}(\mathbb{Q}; \bar{\mathbb{Q}}) &:= \mathcal{D}_{KL}(\mathbb{Q}\bar{\mathbb{Q}} \parallel \mathbb{Q} \otimes \bar{\mathbb{Q}}) \geq \hat{\mathcal{I}}^{(DV)}_{\omega_+}(\mathbb{Q}; \bar{\mathbb{Q}}) \\
&:= \mathbb{E}_{\mathbb{Q}\bar{\mathbb{Q}}}[T_{\omega_+}(q; \bar{q})] - \log \mathbb{E}_{\mathbb{Q}\otimes\bar{\mathbb{Q}}}[e^{T_{\omega_+}(q;\bar{q})}], \\
\mathcal{I}(\mathbb{K}; \bar{\mathbb{Q}}) &:= \mathcal{D}_{KL}(\mathbb{K}\bar{\mathbb{Q}} \parallel \mathbb{K} \otimes \bar{\mathbb{Q}}) \geq \hat{\mathcal{I}}^{(DV)}_{\omega_-}(\mathbb{K}; \bar{\mathbb{Q}}) \\
&:= \mathbb{E}_{\mathbb{K}\bar{\mathbb{Q}}}[T_{\omega_-}(k; \bar{q})] - \log \mathbb{E}_{\mathbb{K}\otimes\bar{\mathbb{Q}}}[e^{T_{\omega_-}(k;\bar{q})}].
\end{aligned}$$

We use MI loss to locally optimize SFM by minimizing $\mathcal{I}(\mathbb{K}; \bar{\mathbb{Q}})$ (i.e., decoupling entangled semantic representations from the overlapping semantic region) and maximizing $\mathcal{I}(\mathbb{Q}; \bar{\mathbb{Q}})$ (i.e., moving disentangled semantic representations into the non-overlapping semantic region away from the majority class). Therefore, MI loss is defined as follows:

$$\mathcal{L}(\hat{\omega}, \hat{\psi}) = \min_{\omega,\psi}[\mathcal{I}(\mathbb{K}; \bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q}; \bar{\mathbb{Q}})].$$

CAG The coupled adversarial generator generates new texts from decoupled semantic representations and the original majority samples. The goal of CAG is to obtain anchor instances with similar surface forms to the input difficult samples, without destroying the original data distribution.

To this end, we make $\mathbb{U}_{\psi_-}$ and $\mathbb{U}_{\psi_+,\sigma,\omega}$, match the prior distributions $\mathbb{N}^-$ and $\mathbb{N}^+$, we introduce two discriminators D_{ϕ_-} and D_{ϕ_+}, each of which is composed of a single hidden layer (ϕ_- and ϕ_+ denote parameters of the majority and minority discriminators). The reconstruction losses of the two decoders, $De_{|\psi_+}$ with parameters | ψ_+ and $De_{|\psi_-}$ with parameters | ψ_-, are denoted as $\mathcal{D}(\mathbb{N}^- \parallel \mathbb{U}_{\psi_-})$ and $\mathcal{D}(\mathbb{N}^+ \parallel \mathbb{U}_{\psi_+,\sigma,\omega})$. $\mathbb{U}_{\psi_-}$ is a marginal

distribution of $\mathbf{Y}^- := \{y_i^- = E_{\psi_-}(x_i^-)|x_i^- \in \mathbf{X}^-\}$, where $\mathbf{Y}^-$ represents the outputs of the majority decoder from the majority input samples $\mathbf{X}^-$ and E_{ψ_-} denotes the encoder-decoder for majority class with parameters $\psi_- \supseteq \{\psi_- \mid \cup \mid \psi_-\}$. Similarly, the minority encoder-decoder is denoted as E_{ψ_+} with its parameters $\psi_+ \supseteq \{\psi_+ \mid \cup \mid \psi_+\}$. The training objectives of the two encoder-decoders in CAG are defined as:

$$\begin{aligned}\mathcal{L}(\hat{\psi}_+, \sigma, \omega, \hat{\phi}_+) &= \min_{\psi_+} \max_{\phi_+} \mathcal{D}_{\phi_+}(\mathbb{N}^+ \parallel \mathbb{U}_{\psi_+,\sigma,\omega}) \\ &= \min_{\psi_+} \max_{\phi_+} \big\{\mathbb{E}_{\mathbb{N}^+}[\log D_{\phi_+}(x^+)] \\ &\quad + \mathbb{E}_{\mathbb{U}}[\log(1 - D_{\phi_+}(E_{\psi_+}(x^+)))]\big\},\end{aligned}$$

$$\begin{aligned}\mathcal{L}(\hat{\psi}_-, \hat{\phi}_-) &= \min_{\psi_-} \max_{\phi_-} \mathcal{D}_{\phi_-}(\mathbb{N}^- \parallel \mathbb{U}_{\psi_-}) \\ &= \min_{\psi_-} \max_{\phi_-} \big\{\mathbb{E}_{\mathbb{N}^-}[\log D_{\phi_-}(x^-)] \\ &\quad + \mathbb{E}_{\mathbb{U}}[\log(1 - D_{\phi_-}(E_{\psi_-}(x^-)))]\big\},\end{aligned}$$

where $\mathbb{E}_\bullet$ estimates the expectation over samples from the $\bullet$ distribution. Both the minority and majority decoders run in an autoregressive way to generate tokens. Taking the generation of an anchor instance as an example, anchor instance is generated as a sequence of l tokens $y_i^+ = (y_{i1}^+, ..., y_{il}^+)$, and $y_{ij}^+ = \arg\max_{\tilde{y}_{ij}^+} p(z_i) \cdot \prod_{j=1}^{l}(\tilde{y}_{ij}^+ | y_{i,<j}^+, z_i)$.

5 Training and Inference

Training Objective In summary, the goal of MISO is to use CAG to generate anchor instance via constructing $\mathbf{Z}$ as disentangled semantic representations by SFM jointly with MI loss, $S_\sigma \circ T_\omega$. The final function is defined as follows:

$$\begin{aligned}\mathcal{L}(\hat{\psi}, \sigma, \omega, \hat{\phi}) &= (1-\alpha)\mathcal{L}(\hat{\omega}, \hat{\psi}) \\ &\quad + \alpha[\mathcal{L}(\hat{\psi}_+, \sigma, \omega, \hat{\phi}_+) + \mathcal{L}(\hat{\psi}_-, \hat{\phi}_-)],\end{aligned}$$

where $\alpha = \frac{1}{\tau}$ (see §4.2 for the definition of τ) is a parameter to control contributions from the MI loss and reconstruction loss.

Training In order to calculate the mutual information, two MINEs need to be pre-trained before training the entire MISO. Furthermore, the warm-up of the minority encoder is a necessary condition for pre-training MINEs, ensuring that the inputs of MINEs are reliable. Therefore, we first freeze SFM and the discriminators to pre-train the minority encoder-decoder. Secondly, we follow the method proposed by Belghazi et al. (2018) to pre-train MINEs. We have found that the training challenge lies in how to train MINEs when SFM is frozen. To solve this, we simulate the output of SFM using a trick. Notably, we concatenate $\mathbf{K}$ and $\mathbf{Q}$ to obtain a set of $2d$-dimensional vectors and feed them into a feedforward neural network to obtain a set of d-dimensional vectors $\tilde{\mathbf{Q}}$. We use $\tilde{\mathbf{Q}}$ as the inputs of the decoders to participate in the pre-training of the encoder-decoder. Finally, we can use $\mathbf{Q}, \tilde{\mathbf{Q}}, \mathbf{K}$ as the inputs to pre-train their MINEs. The whole training process is shown in Algorithm 1 (lines 1-9).

Inference Once MISO is trained, we can use it to generate anchor instances for difficult samples of the minority class. Lines 10-12 in Algorithm 1 demonstrate the inference with MISO. This oversampling of anchor instances will not stop until the two classes are balanced.

Algorithm 1: Training and Inference

Input: $\mathbf{X}^+$: minority examples; $\mathbf{X}^-$: majority examples;
Output: $\mathbf{Y}^+$: anchor instances for difficult samples;
1 Freeze the parameters σ of SFM, ω of two MINEs and ϕ of the discriminators;
2 Pre-train the parameters ψ of the two encoder-decoders by descending their stochastic gradient:$\nabla_\psi[\mathcal{L}(\hat{\psi}_+, \sigma, \omega, \hat{\phi}_+) + \mathcal{L}(\hat{\psi}_-, \hat{\phi}_-)]$;
3 Freeze the parameters σ, ψ and ϕ;
4 Pre-train the parameters ω of MINEs by descending its stochastic gradient:$\nabla_\omega \mathcal{L}(\hat{\omega}, \hat{\psi})$;
5 **for** *number of training iterations* **do**
6 Update the parameters ϕ of the discriminators by ascending their stochastic gradient: $\nabla_\phi[\mathcal{L}(\hat{\psi}_+, \hat{\phi}_+) + \mathcal{L}(\hat{\psi}_-, \hat{\phi}_-)]$;
7 Update the parameters ψ of the two encoder-decoders and σ of SFM by descending their stochastic gradient:$\nabla_{\psi,\sigma}\mathcal{L}(\hat{\psi}, \sigma, \omega, \hat{\phi})$;
8 Update the parameters ω of MINEs by descending its stochastic gradient:$\nabla_\omega \mathcal{L}(\hat{\omega}, \hat{\psi})$;
9 **end**
10 **while** $n^+ + t \leq n^-$ **do**
11 Generating $\mathbf{Y}^+$ with E_+;
12 **end**

6 Experiments

We conducted experiments on several text classification tasks to examine the effectiveness of the proposed MISO against the previous state-of-the-art imbalanced learning methods.

Baselines

- **Focal** and **Dice** loss (re-weighting methods). Lin et al. (2017) and Li et al. (2020) introduce algorithms to learn difficult samples by adjusting the weights of instances.

- **MixText** (data augmentation method). Chen et al. (2020) builds a semi-supervised learning model by interpolating text in the hidden space.

- **ADASYN** (re-sampling method). He et al. (2008) proposes an adaptive synthetic method that generates new examples for each minority instances according to the data distribution.

- **τ-norm** and **cRT** (long-tailed learning methods). Kang et al. (2020) decouple representation learning and classification so as to train the classifier to balance the decision boundary independently.

Note that we have chosen three types of classification models (i.e., TextCNN (Kim , 2014), TextRNN (Liu et al., 2016), and XLNet (Yang et al., 2019)) to be combined with MISO to complete the entire text classification task. The backbone networks of these models are CNN, RNN, and Transformer, respectively. We hence term the combination of them with MISO as M-CNN, M-RNN, and M-XLNet.

Datasets In Table 2, we have selected 6 datasets: 3 imbalanced and 3 balanced datasets. Following by Ger and Klabjan (2019), we changed the balanced datasets into imbalanced datasets by random sampling one of the classes at 1% and 5% in each experiment, which is a common practice in imbalanced learning. Concretely, **Opin-Rank** contains hotel reviews on TripAdvisor and car reviews on Edmunds (Ganesan and Zhai, 2012). **SMS Spam** is created via Short Message Service (SMS) (Peng et al., 2019). **Toutiao** is a Chinese dataset that contains 15 topics (Ouyang et al., 2020). **Yelp.P** contains Yelp reviews about the best restaurants, shopping, nightlife, food, and entertainment (Li et al., 2018). **IMDB** is a movie review dataset (Ger and Klabjan, 2019). **AG_News** consists of news articles from the AG's corpus (Yang et al., 2019). For multi-class datasets (i.e.,Toutiao and AG_News), we treat all data as majority samples, except the data of the selected minority class.

Experiment Settings All experimental results were obtained as the mean of 5-fold cross-validation. We set $b_s = 64$, the learning rate as 1×10^{-4}, $d = 64$, and $h = 8$. We removed stop words by using baidu stop words[1] for Chinese datasets and NLTK 3.5 stop words[2] for English dataset. We selected "Jieba" to do word segmentation on Toutiao dataset[3].

Datasets	C	L	N	P	IR
Opin-Rank	2	144	259,000	42,230	0.16
SMS Spam	2	19	4,601	1,813	0.39
Toutiao	15	18	8,309	68	0.008
Yelp.P	2	153	280,000	2,800	0.01
			280,000	14,000	0.05
IMDB	2	294	25,000	250	0.01
			25,000	1,250	0.05
AG_News	4	91	30,000	300	0.01
			30,000	1,500	0.05

Table 2: Statistics of the used datasets. C: the number of target classes. L: average sentence length. P/N: the number of instances in the minority/majority class. IR: imbalance ratio defined as $\frac{\mathrm{P}}{\mathrm{N}}$.

Evaluation Metrics we adopted F1 metrics (Yan et al., 2019) to evaluate all models.

Results Table 3 summarizes the results of MISO against other methods on each benchmark dataset. MISO achieves the best results on all datasets, suggesting that MISO is consistently effective across different data situations.

Experiment results show that ADASYN, as a widely-used baseline for imbalanced learning, performs not well on imbalanced text classification. The main reason is that the discrete nature of texts results in ADASYN improperly synthesizing data that don't exist in the real world. This destroys the distribution of texts to some extent. Such a problem also appears in MixText. In contrast, MISO leverages CAG to keep the consistency between the new distribution and the original distribution.

Focal and Dice set larger learning weights for difficult samples. This is feasible when minority data is sufficient, and vice versa in Toutiao, IMDB (1%) and AG_News (1%) datasets. Since MISO supplies anchor instances for the minority class, an average of 3.5% improvement can still be obtained in the case of data sparseness.

Following by Kang et al. (2020), we kept the backbone network (i.e., representation learning) frozen, and fine-tuned classifiers by class-balanced sampling (cRT) or decision boundary rectifying (τ-norm). Neither of them considered the impact of difficult samples on searching clear decision boundaries. In contrast, MISO outperforms them by an average of 2.7%. This explicitly illustrates the necessity of re-embedding difficult samples.

[1] http://www.baiduguide.com/baidu-stopwords/
[2] https://www.nltk.org/
[3] http://pypi.python.org/pypi/jieba/

Methods	Opin-Rank	Toutiao	SMS Spam	Yelp.P (1%)	IMDB (1%)	AG_News (1%)	Yelp.P (5%)	IMDB (5%)	AG_News (5%)
TextCNN	92.0	81.9	80.1	75.8	50.1	46.2	82.0	68.1	65.1
TextRNN	92.1	82.0	80.3	75.7	50.1	46.4	82.4	68.6	64.8
XLNet	92.7 •	82.2 •	81.9 •	76.9 •	51.7 •	47.6 •	83.2 •	69.1 •	65.7 •
ADASYN	92.9	83.4	82.2	77.1	53.9	48.7	83.7	74.0	66.9
MixText	94.9	84.9	86.1	77.6	56.9	51.5	84.4	76.1	66.9
Focal	94.9	83.9	86.6	77.5	54.1	49.3	84.8	76.0	67.7
Dice	94.7	84.0	86.2	77.5	54.9	49.3	85.6	76.3	67.6
cRT	95.1	85.5 ○	87.2 ○	78.1 ○	59.5	50.9	85.9	76.5 ○	68.2
τ-norm	95.1 ○	85.3	87.1	78.0	59.6 ○	52.5 ○	86.0 ○	76.4	68.4 ○
CAG (Ours)	94.1 (-1.0)	84.9 (-0.6)	84.8 (-2.4)	78.0 (-0.1)	57.0 (-2.6)	52.3 (-0.2)	85.3 (-0.7)	75.0 (-1.5)	67.6 (-0.8)
M-CNN (Ours)	**96.0** (+0.9)	87.2	**89.0** (+1.8)	81.4	62.2	55.9	87.2	**78.4** (+1.9)	71.9
M-RNN (Ours)	95.4	86.6	89.0	81.9	62.3	55.5	87.0	78.1	72.2
M-XLNet (Ours)	95.6	**88.4** (+2.9)	88.0	**82.5** (+4.4)	**63.4** (+3.8)	**56.4** (+3.9)	**87.3** (+1.3)	77.9	**72.2** (+3.8)

Table 3: Experiment results of imbalanced text classification. •: the best performance of classifiers. ○: the best performance of baselines with different classifiers. Bold: the best performance of MISO with different classifiers. Underscore: the best performance of the oversampling model with different classifiers.

Datasets	Perc. (%) of DS	F1-score (%) of DS | Non-DS		
		TextCNN → M-CNN (Ours)	TextRNN → M-RNN (Ours)	XLNet → M-XLNet (Ours)
Opin-Rank	15.3 → 8.3 (-7.0)	54.9 | 98.7 → 55.8 | 99.0	54.9 | 98.8 → 55.7 | 99.7	56.2 | 99.3 → 56.3 | 99.4
SMS Spam	42.1 → 22.0 (-22.1)	61.9 | 93.2 → 62.1 | 93.4	61.9 | 93.6 → 62.5 | 94.1	64.3 | 94.6 → 65.9 | 94.8
Toutiao	36.6 → 17.7 (-18.9)	59.0 | 95.1 → 59.6 | 95.4	59.0 | 95.2 → 59.4 | 95.4	59.0 | 95.6 → 61.1 | 95.9
Yelp.P (1%)	28.2 → 20.1 (-8.1)	29.1 | 94.2 → 30.0 | 94.9	29.0 | 94.0 → 29.4 | 94.4	32.6 | 94.3 → 33.5 | 94.9
Yelp.P (5%)	22.4 → 13.8 (-8.6)	41.9 | 93.6 → 41.9 | 94.2	43.7 | 93.5 → 44.0 | 94.1	46.4 | 93.8 → 46.5 | 93.9
IMDB (1%)	49.3 → 31.4 (-17.9)	16.8 | 82.4 → 17.0 | 83.1	16.8 | 82.4 → 17.5 | 82.6	18.8 | 83.6 → 19.1 | 83.9
IMDB (5%)	30.1 → 15.6 (-14.5)	23.6 | 87.3 → 24.0 | 88.1	21.9 | 87.9 → 22.9 | 88.6	26.9 | 87.3 → 27.1 | 87.3
AG_News (1%)	48.5 → 33.5 (-15.0)	15.5 | 75.1 → 16.1 | 75.4	15.4 | 75.5 → 16.2 | 76.0	17.5 | 75.9 → 17.6 | 76.1
AG_News (5%)	33.2 → 21.7 (-11.5)	27.7 | 84.0 → 28.4 | 84.3	24.6 | 84.7 → 25.1 | 84.9	27.7 | 83.8 → 27.9 | 83.9

Table 4: Statistical analysis on the impact of difficult sample re-embedding on each dataset. DS: difficult samples. Non-DS: non-difficult samples. Perc.: percentages.

In addition, MISO enables models based on CNN or RNN, without pre-training, to outperform XLNet in Opin-Rank, SMS Spam, and IMDB datasets, thus saving time and space for training.

7 Analysis

We carried out the statistical and empirical analysis to the superiority of MISO for re-embedding difficult samples.

7.1 Difficult Sample Re-embedding

We counted the number γ of majority samples in the k-nearest neighbors of each minority sample. If γ >0, the corresponding minority sample is considered as a difficult sample.

Results Table 4 shows statistics on the change of difficult samples before and after MISO is used on different datasets. The average decrease in the percentage of difficult samples on all datasets is 13.7% after re-embedding difficult samples. This illustrates that MISO can effectively transform the semantic representation of entangled difficult sample into a non-difficult version. Surprisingly, the F1-score of difficult samples and non-difficult samples do not decrease, which suggests that re-embedding difficult samples and generating anchor instances do not make classifiers lose their ability to classify non-difficult samples. Intuitively, while the classification performances over difficult samples and non-difficult samples can be maintained, as some difficult samples become non-difficult samples, the overall classification performance will inevitably be improved.

7.2 Ablation Study

As mentioned above, decoupling entangled semantic representation of difficult sample from the majority class is achieved by SFM jointly with MI loss. Therefore, in order to verify the effectiveness of this method, we specially conducted ablation experiments: only using CAG to conduct the above comparative experiment (see Table 3) and analysis of difficult sample re-embedding (see Table 5).

Results Compared with the state-of-the-art methods, CAG has an average performance drop of 1.1%. This indicates that SFM constrained by the MI loss effectively improves the overall performance of classifiers with an average of 3.8%. Merely using CAG to generate new texts is actually a sample-balanced sampling, and shares with cRT in that they all ignore the decoupling of difficult samples. See Table 5, the percentages of difficult samples only decreases by an average of 1.1%. In Yelp.P (5%) and AG_News (1%), the per-

Datasets	Perc. (%) of DS	F1-score (%) of DS | Non-DS		
		TextCNN → M-CNN (Ours)	TextRNN → M-RNN (Ours)	XLNet → M-XLNet (Ours)
Opin-Rank	15.3 → 14.2 (-1.1)	54.9 | 98.7 → 54.9 | 99.0	54.9 | 98.8 → 54.5 | 98.7 ⋄	56.2 | 99.3 → 56.1 | 98.9 ⋄
SMS Spam	42.1 → 40.5 (-1.6)	61.9 | 93.2 → 62.0 | 93.3	61.9 | 93.6 → 62.2 | 93.8	64.3 | 94.6 → 64.5 | 94.3 ▷
Toutiao	36.6 → 34.2 (-2.2)	59.0 | 95.1 → 58.5 | 94.9 ⋄	59.0 | 95.2 → 59.2 | 95.4	59.0 | 95.6 → 59.3 | 95.2 ▷
Yelp.P (1%)	28.2 → 27.5 (-0.7)	29.1 | 94.2 → 29.0 | 94.2 ◁	29.0 | 94.0 → 28.8 | 94.4 ◁	32.6 | 94.3 → 33.0 | 94.2 ▷
Yelp.P (5%)	22.4 → 22.6 (+0.2) †	41.9 | 93.6 → 41.6 | 93.9 ◁	43.7 | 93.5 → 43.6 | 93.1 ⋄	46.4 | 93.8 → 45.9 | 94.3 ◁
IMDB (1%)	49.3 → 47.3 (-2.0)	16.8 | 82.4 → 17.0 | 82.7	16.8 | 82.4 → 16.5 | 82.8 ◁	18.8 | 83.6 → 18.9 | 83.9
IMDB (5%)	30.1 → 29.5 (-0.6)	23.6 | 87.3 → 23.4 | 87.2 ⋄	21.9 | 87.9 → 21.7 | 87.5 ⋄	26.9 | 87.3 → 27.0 | 87.3
AG_News (1%)	48.5 → 49.1 (+0.6) †	15.5 | 75.1 → 15.7 | 74.6 ▷	15.4 | 75.5 → 15.8 | 75.5	17.5 | 75.9 → 17.0 | 75.9 ◁
AG_News (5%)	33.2 → 30.9 (-2.3)	27.7 | 84.0 → 27.2 | 83.6 ⋄	24.6 | 84.7 → 24.5 | 84.8 ◁	27.7 | 83.8 → 27.9 | 83.8

Table 5: The ablation experiment of MISO on each dataset. DS: difficult samples. Non-DS: non-difficult samples. Perc.: percentages. †: ascent in the number of difficult samples. ⋄: descent in F1-score of all samples. ◁: descent in F1-score of difficult samples. ▷: descent in F1-score of non-difficult samples.

Dataset	Difficult samples	Anchor instances by MISO
SMS Spam	**Just text ok to us** and we'll credit your account.	**Just text ok to us** that guarantee your bonus.
Opin-Rank	The **sonata** has a very smooth ride and great pick.	The **sonata** has ample acceleration because of our window setting.
Toutiao	九尾**狐狸**为何要附体在**苏妲己**身上? 只因妲己的一项特质无以伦比.	玉面**狐狸**附体**苏妲己**, 五年后救其一命报恩.
Yelp.P	1.**My parents didn't want to** go back to beautiful Miami. 2.When the **guy** came to the door, he said it was late.	1.To my disappointment, **my parents didn't want to** help me. 2.Bad, these **guys** are very slow. 3.These **guys** were rude and I really have a disappointing meal.
IMDB	A **incredible story** about a man who wants to figure out what really happened ...	1.One of wood 's oscars ! **incredible story** ! ! 2.What you **incredible story** ! brilliant!
AG_News	The **manager** said he left **North London** because he can not control recruitment.	The **manager** of the **North London** football club will be banned for the next seven years.
CAG (Ours)	1. I was **worst** ridiculous by **worst** restaurant. 2. The complimentary **worst** and oil was **worst**. 3. I angry their thing **worst**. 4. Its a bit session for us but **worst** once in a while.	

Table 6: Examples of anchor instances.

centages of difficult samples even increase by 0.2% and 0.6%, respectively, so that the newly added difficult samples will inevitably make the classification boundary more difficult to capture. In addition, based on the experimental results of Kang et al. (2020), the effect of class-balance sampling and decision boundary rectifying are slightly better than that of sample-balance sampling, which is the main reason that CAG is more inefficient than the state-of-the-art methods. It is important to note that CAG also degrades the classification performance, especially for non-difficult samples in SMS Spam, Toutiao, Yelp.P (1%) and AG_News (1%) datasets (see Table 5), however, this issue does not appear in MISO (see Table 4).

7.3 Case Study

We looked into our data to investigate how MISO generate anchor instances.

Results In Table 6, because of the imbalance problem, when all tokens in the spam SMS "*Just text ok to us and we will credit your account*" often appear in non-spam SMS with a high frequency, the classifier is misguided to categorize this sentence as non-spam SMS. To solve this issue, MISO generates anchor instance by adding tokens such as "*guarantee*" and "*bonus*". The anchor instance makes the backbone network re-embed this difficult sample in a non-overlapping form, which is more likely to be correctly classified as a spam SMS. A more interesting example appears in Yelp.P, where MISO seems to learn the semantic entailment of the original difficult sample, that is, "*it was late*" entails "*guys are slow*". These examples suggest that MISO is able to learn the underlying meaning of difficult samples and generate new samples that preserve the original meaning.

We also conducted the ablation experiment of CAG for the case study. Due to the space limitations, we only show the Yelp.P (1%) example in Table 6. The repeated token "*worst*" in instances generated by CAG are usually meaningless. This reflects that without SFM and MI loss, CAG only learn the extremely limited semantics.

8 Conclusion

In this paper, we have presented an effective mutual information-constrained oversampling strategy, which re-embed difficult samples via a safe and robust method. Our method makes the traditional

text classification still feasible when dealing with imbalanced data in the real world. In future work, we will try to design a more effective backbone network for re-embedding difficult samples.

9 Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (61972455) and the Joint Project of Bayescom. Xiaowang Zhang is supported by the program of Peiyang Young Scholars in Tianjin University (2019XRX-0032).

References

Mohamed I. Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *Proceedings of the 35th InternMational Conference on Machine Learning (ICML)*, pages 531–540.

Jason Brownlee. 2019. A gentle introduction to imbalanced classification. *Machine Learning Mastery*.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Proceedings of the 32th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1565–1576.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2147–2157.

Monroe D. Donsker and Varadhan. 1975. Asymptotic evaluation of certain markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47.

Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645.

Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from Imbalanced Data Sets*. Springer, Berlin, DE.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 663–670.

Kavita Ganesan and Chengxiang Zhai. 2012. Opinion-based entity ranking. *Information Retrieval*, 15(2):116–150.

Yang Gao, Yifan Li, Yu Lin, Charu C. Aggarwal, and Latifur Khan. 2020. RiSAWOZ: A large-scale multi-domain wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940.

Stephanie Ger and Diego Klabjan. 2019. Autoencoders and generative adversarial networks for anomaly detection for sequences. arXiv:1901.02514.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587.

Anna Glazkova. 2020. A comparison of synthetic oversampling methods for multi-class text classification. arXiv:2008.04636.

Hidalgo J.M. Gomez, Manuel J.M. López, and Enrique P. Sanz. 2000. Combining text and heuristics for cost-sensitive spam filtering. In *Proceedings of the 4th Conference on on Computational Natural Language Learning (CoNLL)*, pages 99–102.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5141–5148.

Hui Han, Wenyuan Wang, and Binghuan Mao. 2005. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Proceedings of the 17th International Conference on Intelligent Computing (ICIC)*, pages 878–887.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 7th International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, pages 26–30.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1865–1874.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 465–476.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the 30th IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Shigang Liu, Yu Wang, Jun Zhang, Chao Chen, and Yang Xiang. 2017. Addressing the class imbalance problem in twitter spam detection using ensemble learning. *Computers & Security*, 69:35–49.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-Task learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2873–2879.

Ryan McBride, Ke Wang, Zhouyang Ren, and Wenyuan Li. 2019. Cost-sensitive learning to rank. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence (AAAI)*, pages 4570–4577.

Weili Nie, Nina Narodytska, and Ankit Patel. 2019. RelGAN: Relational generative adversarial networks for text generation. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, pages 1–20.

Kemal Oksuz, Baris Can Cam, Sinan Kalkan, and Emre Akbas. 2019. Imbalance problems in object detection: A review. arXiv:1503.06733.

Wentao Ouyang, Xiuwu Zhang, Lei Zhao, Jinmei Luo, Yu Zhang, Heng Zou, Zhaojie Liu, and Yanlong Du. 2020. Minet: Mixed interest network for cross-domain click-through rate prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 2669–2676.

Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. 2019. Libra R-CNN: Towards balanced learning for object detection. In *Proceedings of the 32th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 821–830.

Minlong Peng, Qi Zhang, Xiaoyu Xing, Tao Gui, Xuanjing Huang, Yugang Jiang, Keyu Ding, and Zhigang Chen. 2019. Trainable undersampling for class-imbalance learning. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence (AAAI)*, pages 4707–4714.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A large-scale multi-domain wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 91–99.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. arXiv:2010.04592.

Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. 2017. A fast R-CNN: Hard positive generation via adversary for object detection. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2606–2615.

Fangzhao Wu, Chuhan Wu, and Junxin Liu. 2018. Imbalanced sentiment classification with multi-task learning. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1631–1634.

Wenshuo Yang, Jiyi Li, Fumiyo Fukumoto, and Yanming Ye. 2020. HSCNN: A hybrid-siamese convolutional neural network for extremely imbalanced multi-label text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6716–6722.

Yuguang Yan, Mingkui Tan, Yanwu Xu, Jiezhang Cao, Michael Ng, Huaqing Min, and Qingyao Wu. 2019. Oversampling for imbalanced data via optimal transport. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence (AAAI)*, pages 5605–5612.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNET: Generalized autoregressive pretraining for language understanding. In *Proceedings of the 32th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 5754–5764.

Appendix

Background

Mutual Information In probability theory and information theory, mutual information (MI) measures the interdependence between two random distributions. It can be used for estimating the similarity between the joint distribution and the product of marginal distributions.

For convenience, we abbreviate $P_{x\sim\mathbb{X}}(x)$ as $P(\mathbb{X})$, where $\mathbb{X}$ denotes any distribution of $x \in X$. The entropy of distribution $\mathbb{X}$ can be defined as

$$\mathcal{H}(\mathbb{X}) = -\sum_{X} P(\mathbb{X}) \log P(\mathbb{X}). \tag{1}$$

Given two probability distributions $\mathbb{A}$ and $\mathbb{B}$ taking values from finite sets A and B respectively, and $x \in A, y \in B$, the conditional entropy of A given B can be defined as

$$\mathcal{H}(\mathbb{A} \mid \mathbb{B}) = -\sum_{A}\sum_{B} P(\mathbb{A}\mathbb{B}) \log P(\mathbb{B} \mid \mathbb{A}). \tag{2}$$

We formalize MI from the perspective of probability theory. We define the joint distribution of A and B is $P_{\mathbb{A}\mathbb{B}}(x, y)$. Then, the discrete probability version of the mutual information can be formalized as

$$\mathcal{I}(\mathbb{A};\mathbb{B}) = \sum_{B}\sum_{A} P(\mathbb{A}\mathbb{B}) \log \frac{P(\mathbb{A}\mathbb{B})}{P(\mathbb{A})P(\mathbb{B})}. \tag{3}$$

Specifically, the mutual information between $\mathbb{A}$ and $\mathbb{B}$ is the reduction in the uncertainty of $\mathbb{A}$ due to the knowledge of $\mathbb{B}$ (or vice versa). Therefore, it can be defined as

$$\begin{aligned}\mathcal{I}(\mathbb{A};\mathbb{B}) =& \mathcal{H}(\mathbb{A}) - \mathcal{H}(\mathbb{A} \mid \mathbb{B}) \\ =& \mathcal{H}(\mathbb{B}) - \mathcal{H}(\mathbb{B} \mid \mathbb{A}).\end{aligned}$$

According to the above definition, the mutual information satisfies the following properties:

- Non-negativity (i.e., $\mathcal{I}(\mathbb{A};\mathbb{B}) \geq 0$);
- Symmetry (i.e., $\mathcal{I}(\mathbb{A};\mathbb{B}) = \mathcal{I}(\mathbb{B};\mathbb{A})$).

In addition, its extremum property is:

$$\mathcal{I}(\mathbb{A};\mathbb{B}) \leq -\sum_{A} P(\mathbb{A}) \log P(\mathbb{A}) \leq \log |A|.$$

where $|A|$ is the size of the set A.

Kullback-Leibler (KL) Divergence KL divergence can be viewed as a measure of "distance" or "dissimilarity" between distributions $\mathbb{A}$ and $\mathbb{B}$, defined over a common alphabet X and written as $\mathcal{D}(\mathbb{A}\|\mathbb{B})$. It measures the inefficiency of mistakenly assuming that the distribution of a source is $\mathbb{B}$ when the true distribution is $\mathbb{A}$. Similarly, the definition of KL divergence can be defined by:

$$\mathcal{D}(\mathbb{A}\|\mathbb{B}) = \sum_{X} P(\mathbb{A}) \log \frac{P(\mathbb{B})}{P(\mathbb{A})}.$$

Then, KL divergence satisfies:

- Non-negativity (i.e., $\mathcal{D}(\mathbb{A}\|\mathbb{B}) \geq 0$);
- Asymmetry (i.e., $\mathcal{D}(\mathbb{A}\|\mathbb{B}) \neq \mathcal{D}(\mathbb{B}\|\mathbb{A})$.

Variational Distance The variational distance (also known as the $\mathcal{L}_1-$ distance) between two distributions $\mathbb{A}$ and $\mathbb{B}$ with X is defined by

$$\|\mathbb{A} - \mathbb{B}\| = \sum_{x} |P(\mathbb{A}) - P(\mathbb{B})|.$$

Thus, it satisfies:

- Non-negativity (i.e., $\|\mathbb{A} - \mathbb{B}\| \geq 0$);
- Symmetry (i.e., $\|\mathbb{A} - \mathbb{B}\| = \|\mathbb{B} - \mathbb{A}\|$.

In addition, variational distance and KL divergence satisfy:

$$\mathcal{D}(\mathbb{A}\|\mathbb{B}) \geq \frac{\log_2(e)}{2}\|P(\mathbb{A}) - P(\mathbb{B})\|^2, \tag{4}$$

which is referred to as Pinsker's inequality.

Learning the (Near-)optimal Boundary of the Minority Class

Theorem 1 *Suppose $\mathcal{H}(\mathbb{Q}) \leq \mathcal{H}(\mathbb{K})$ and $\bar{\mathbb{Q}}$ is completely determined by $\mathbb{K}$ and $\mathbb{Q}$ (i.e., $P(\mathbb{K}|\bar{\mathbb{Q}}) + P(\mathbb{Q}|\bar{\mathbb{Q}}) \geq 1$), then*

$$\mathcal{L}(D) \leq \min_{x\in\bar{\mathbf{Q}}}[\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})].$$

Furthermore, $\min_{x\in\bar{\mathbf{Q}}}[\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})] = -\mathcal{H}(\mathbb{Q})$, iff $P(\mathbb{Q}|\bar{\mathbb{Q}}) = 1$ and $P(\mathbb{K}|\bar{\mathbb{Q}}) = P(\mathbb{K})$.

Proof. We map $\mathbb{Q}, \mathbb{K}$ to $\bar{\mathbb{Q}}$. Finding the optimal discriminator D^* by fixing generator G, so that $\mathcal{L}(D)$ reaches maximum:

$$D^* = \arg\max_{D} \mathcal{L}(D).$$

By calculation,

$$\begin{aligned}\mathcal{L}(D) =& \mathop{\mathbb{E}}_{(\mathbb{Q}|\bar{\mathbb{Q}})}[P(\bar{\mathbb{Q}})\log(D(x))] \\ &+ \mathop{\mathbb{E}}_{(\mathbb{K}|\bar{\mathbb{Q}})}[P(\bar{\mathbb{Q}})\log(1-D(x))] \\ &= \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}|\bar{\mathbb{Q}})P(\bar{\mathbb{Q}})\log(D(x))\mathrm{d}x \\ &+ \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}|\bar{\mathbb{Q}})P(\bar{\mathbb{Q}})\log(1-D(x))\mathrm{d}x \\ &= \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}})\log(D(x))\mathrm{d}x \\ &+ \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}})\log(1-D(x))\mathrm{d}x \\ &= \sum_{x\in\bar{\mathbf{Q}}} [P(\mathbb{Q}\bar{\mathbb{Q}})\log(D(x)) \\ &+ P(\mathbb{K}\bar{\mathbb{Q}})\log(1-D(x))]\mathrm{d}x.\end{aligned}$$

Let $\Delta := P(\mathbb{Q}\bar{\mathbb{Q}})\log(D(x)) + P(\mathbb{K}\bar{\mathbb{Q}})\log(1-D(x))$, For $x \in \bar{\mathbf{Q}}$, the maximization of $\mathcal{L}(D)$ is equivalent to the maximization of Δ. Derivation of Δ can be obtained by:

$$\frac{\mathrm{d}}{\mathrm{d}D}\Delta = \frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{D} - \frac{P(\mathbb{K}\bar{\mathbb{Q}})}{1-D}. \tag{5}$$

Observe

$$\frac{\mathrm{d}^2\Delta}{\mathrm{d}D} = -\big[\frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{D^2} - \frac{P(\mathbb{K}\bar{\mathbb{Q}})}{(1-D)^2}\big] < 0.$$

Therefore, Δ is a concave with respect to the discriminator D. When $\frac{\mathrm{d}}{\mathrm{d}D}\Delta = 0$, Δ is the maximization. In other words,

$$\frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{D} - \frac{P(\mathbb{K}\bar{\mathbb{Q}})}{1-D} = 0. \tag{6}$$

Then the optimal D^* is computed as follows:

$$D^* = \frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\mathbb{K}\bar{\mathbb{Q}}) + P(\mathbb{Q}\bar{\mathbb{Q}})},$$

Furthermore, $\mathcal{L}(D)$ can be formulated as,

$$\begin{aligned}\mathcal{L}(D) =& \mathop{\mathbb{E}}_{(\mathbb{Q}|\bar{\mathbb{Q}})}\left[P(\bar{\mathbb{Q}})\log\frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})}\right] \\ &- \mathop{\mathbb{E}}_{(\mathbb{K}|\bar{\mathbb{Q}})}\left[P(\bar{\mathbb{Q}})\log\big(1-D\big(\frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})}\big)\big)\right] \\ &= \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}})\log\big(\frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})}\big)\mathrm{d}x \\ &+ \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}})\log\big(\frac{P(\mathbb{K}\bar{\mathbb{Q}})}{P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})}\big)\mathrm{d}x.\end{aligned} \tag{7}$$

Based on (7) and the definition of the mutual information (3), we can derive

$$\begin{aligned}&\mathcal{L}(D) - [\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})] \\ =& \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}})\log\Bigg[\frac{P(\mathbb{K}\bar{\mathbb{Q}})}{P(\mathbb{K})P(\bar{\mathbb{Q}})} \\ &\times \frac{P(\mathbb{K})P(\bar{\mathbb{Q}})}{P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})}\Bigg] \\ &+ \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}})\log\Bigg[\frac{P(\mathbb{Q})P(\bar{\mathbb{Q}})}{P(\mathbb{Q}\bar{\mathbb{Q}})} \\ &\times \frac{P^2(\mathbb{Q}\bar{\mathbb{Q}})}{(P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}}))P(\mathbb{Q})P(\bar{\mathbb{Q}})}\Bigg] \\ &- \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}})\log\frac{P(\mathbb{K}\bar{\mathbb{Q}})}{P(\mathbb{K})P(\bar{\mathbb{Q}})} \\ &- \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}})\log\frac{P(\mathbb{Q})P(\bar{\mathbb{Q}})}{P(\mathbb{Q}\bar{\mathbb{Q}})} \\ =& \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}})\log\frac{P(\mathbb{K})P(\bar{\mathbb{Q}})}{P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})} \\ &+ \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \\ &\times \log\frac{P^2(\mathbb{Q}\bar{\mathbb{Q}})}{(P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}}))P(\mathbb{Q})P(\bar{\mathbb{Q}})} \\ =& \sum_{x\in\bar{\mathbf{Q}}} \Big\{P(\mathbb{K}\bar{\mathbb{Q}})\log[P(\mathbb{K})P(\bar{\mathbb{Q}})] \\ &- P(\mathbb{K}\bar{\mathbb{Q}})\log[P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})] \\ &+ 2P(\mathbb{Q}\bar{\mathbb{Q}})\log P(\mathbb{Q}\bar{\mathbb{Q}}) \\ &- P(\mathbb{Q}\bar{\mathbb{Q}})\log[P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})] \\ &- P(\mathbb{Q}\bar{\mathbb{Q}})\log[P(\mathbb{Q})P(\bar{\mathbb{Q}})]\Big\} \\ =& \sum_{x\in\bar{\mathbf{Q}}} \Big\{P(\mathbb{K}\bar{\mathbb{Q}})\log P(\mathbb{K}) - P(\mathbb{Q}\bar{\mathbb{Q}})\log P(\mathbb{Q}) \\ &- 2P(\mathbb{Q}\bar{\mathbb{Q}})\log P(\bar{\mathbb{Q}}) + 2P(\mathbb{Q}\bar{\mathbb{Q}})\log P(\mathbb{Q}\bar{\mathbb{Q}}) \\ &- [P(\mathbb{Q}\bar{\mathbb{Q}})+P(\mathbb{K}\bar{\mathbb{Q}})]\log[P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})] \\ &+ P(\mathbb{K}\bar{\mathbb{Q}})\log P(\bar{\mathbb{Q}}) + P(\mathbb{Q}\bar{\mathbb{Q}})\log P(\bar{\mathbb{Q}})\Big\} \\ =& \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}})\log P(\mathbb{K}) - \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}})\log P(\mathbb{Q}) \\ &+ 2\sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}})\log\frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\bar{\mathbb{Q}})} \\ &- \sum_{x\in\bar{\mathbf{Q}}} [P(\mathbb{Q}\bar{\mathbb{Q}})+P(\mathbb{K}\bar{\mathbb{Q}})] \\ &\times \log\frac{P(\mathbb{K}\bar{\mathbb{Q}})+P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\bar{\mathbb{Q}})} \\ =& J(\mathbf{I}) + J(\mathbf{II})\end{aligned} \tag{8}$$

where

$$
\begin{aligned}
J(\mathbf{I}) = &\sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}}) \log P(\mathbb{K}) \\
&- \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log P(\mathbb{Q}) \\
&+ 2\sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log \frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\bar{\mathbb{Q}})};
\end{aligned} \tag{9}
$$

For $J(\mathbf{I})$, due to $\mathcal{H}(\mathbb{Q}) \leq \mathcal{H}(\mathbb{K})$, we have

$$
\begin{aligned}
J(\mathbf{I}) = &\sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}}) \log P(\mathbb{K}) \\
&- \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log P(\mathbb{Q}) \\
&+ 2\sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log \frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\bar{\mathbb{Q}})} \\
= &\sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}}) \log P(\mathbb{K}) \\
&+ \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log \frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\bar{\mathbb{Q}})P(\mathbb{Q})} \\
&+ \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log P(\mathbb{Q}|\bar{\mathbb{Q}}) \\
\leq &\sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}) \log P(\mathbb{K}) \\
&- \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}) \log P(\mathbb{Q}) \\
&+ 2\sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log P(\mathbb{Q}|\bar{\mathbb{Q}}) \\
= &\mathcal{H}(\mathbb{Q}) - \mathcal{H}(\mathbb{K}) - \mathcal{H}(\mathbb{Q}|\bar{\mathbb{Q}}) \\
\leq &0
\end{aligned} \tag{10}
$$

For $J(\mathbf{II})$, because $P(\mathbb{K}|\bar{\mathbb{Q}}) + P(\mathbb{Q}|\bar{\mathbb{Q}}) \geq 1$, we can obtain

$$
\begin{aligned}
J() = &-\sum_{x\in\bar{\mathbf{Q}}} \Big\{ [P(\mathbb{Q}\bar{\mathbb{Q}}) + P(\mathbb{K}\bar{\mathbb{Q}})] \\
&\cdot \log \frac{P(\mathbb{K}\bar{\mathbb{Q}}) + P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\bar{\mathbb{Q}})} \Big\} \\
= &-\sum_{x\in\bar{\mathbf{Q}}} \Big\{ [P(\mathbb{Q}\bar{\mathbb{Q}}) + P(\mathbb{K}\bar{\mathbb{Q}})] \\
&\cdot \log[P(\mathbb{K}|\bar{\mathbb{Q}}) + P(\mathbb{Q}|\bar{\mathbb{Q}})] \Big\} \\
\leq &0
\end{aligned} \tag{11}
$$

Furthermore, according to $J(\mathbf{I}), J(\mathbf{II})$, then $\mathcal{L}(D) - [\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})] \leq 0$, that is

$$
\mathcal{L}(D) \leq \mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}}).
$$

Firstly, for $\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}})$, we can obtain:

$$
\begin{aligned}
\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) &= \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}\bar{\mathbb{Q}}) \log \frac{P(\mathbb{K}\bar{\mathbb{Q}})}{P(\mathbb{K})P(\bar{\mathbb{Q}})} \\
&= \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{K}|\bar{\mathbb{Q}})P(\bar{\mathbb{Q}}) \log \frac{P(\mathbb{K}|\bar{\mathbb{Q}})}{P(\mathbb{K})}.
\end{aligned} \tag{12}
$$

Therefore, $\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) = \mathcal{I}(P(\bar{\mathbb{Q}}), P(\mathbb{K}|\bar{\mathbb{Q}}))$.

In addition, since $\mathcal{I}(P(\bar{\mathbb{Q}}), P(\mathbb{K}|\bar{\mathbb{Q}}))$ is the convex function of $P(\mathbb{K}|\bar{\mathbb{Q}})$, then there exists the only $p^*(\mathbb{K}|\bar{\mathbb{Q}})$ defined on $[0,1]$, which satisfies $p^*(\mathbb{K}|\bar{\mathbb{Q}}) = \arg\min \mathcal{I}(P(\bar{\mathbb{Q}}), P(\mathbb{K}|\bar{\mathbb{Q}}))$

According to the properties of convex functions: when $P(\mathbb{K}|\bar{\mathbb{Q}}) = P(\mathbb{K})$, $\mathcal{I}(P(\bar{\mathbb{Q}}), P(\mathbb{K}|\bar{\mathbb{Q}}))$ is minimum, and $\min(\mathbb{K}, \bar{\mathbb{Q}}) = 0$.

Secondly, for $\mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})$, from the logarithmic sum inequality:

$$
\begin{aligned}
\mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}}) &= \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log \frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\mathbb{Q})P(\bar{\mathbb{Q}})} \\
&\geq \sum_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}}) \log \frac{\sum\limits_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}})}{\sum\limits_{x\in\bar{\mathbf{Q}}} [P(\mathbb{Q})P(\bar{\mathbb{Q}})]} \\
&= \mathcal{H}(\mathbb{Q}).
\end{aligned} \tag{13}
$$

With equality holding iff:

$$
\begin{aligned}
&\frac{P(\mathbb{Q}\bar{\mathbb{Q}})}{P(\mathbb{Q})P(\bar{\mathbb{Q}})} = \frac{\sum\limits_{x\in\bar{\mathbf{Q}}} P(\mathbb{Q}\bar{\mathbb{Q}})}{\sum\limits_{x\in\bar{\mathbf{Q}}} [P(\mathbb{Q})P(\bar{\mathbb{Q}})]}; \\
&i.e.\ P(\mathbb{Q}|\bar{\mathbb{Q}}) = 1.
\end{aligned} \tag{14}
$$

Combined $\mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})$ and $\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}})$, when $P(\mathbb{K}|\bar{\mathbb{Q}}) = P(\mathbb{K}), P(\mathbb{Q}|\bar{\mathbb{Q}}) = 1$, $P(\mathbb{K}|\bar{\mathbb{Q}}) + P(\mathbb{Q}|\bar{\mathbb{Q}}) \geq 1$, then, $\min\{\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})\} = -\mathcal{H}(\mathbb{Q})$.

Finally, as $\mathcal{L}(D) \leq \mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})$, we have

$$
\mathcal{L}(D) \leq \min\{\mathcal{I}(\mathbb{K};\bar{\mathbb{Q}}) - \mathcal{I}(\mathbb{Q};\bar{\mathbb{Q}})\} = -\mathcal{H}(\mathbb{Q}).
$$

Theorem 1 is therefore established. □

Distribution Consistency

Theorem 2 *Under the same conditions as Theorem 1, the distribution captured by MISO is consistent with the original distribution of the minority class. Therefore,it is represented as follows:* $\|P(\mathbb{Q}) - P(\bar{\mathbb{Q}})\| = 0$ *as* $P(\mathbb{K}|\bar{\mathbb{Q}}) \to P(\mathbb{K})$ *and* $P(\mathbb{Q}|\bar{\mathbb{Q}}) \to 1$.

Proof. Firstly, we can obtain that minority and majority samples are mutually exclusive and independent of each other by combining the definition of conditional probability and classification problem, that is,

$$P(\mathbb{Q}\mathbb{K}) = P(\mathbb{Q})P(\mathbb{K}),$$

then we have

$$P(\mathbb{Q}|\mathbb{K}) = \frac{P(\mathbb{Q}\mathbb{K})}{P(\mathbb{K})} = P(\mathbb{Q}). \tag{15}$$

Furthermore, by the fact that **conditioning never decreases divergence**,we have

$$D(\mathbb{Q}\|\bar{\mathbb{Q}}) \leq D(\mathbb{Q}\|\bar{\mathbb{Q}}|\mathbb{K}),$$

that is, $\forall x, y, z \in \bar{\mathbf{Q}}$, there is

$$\begin{aligned} &\sum_{x,y} P(\mathbb{Q}) \log \frac{P(\mathbb{Q})}{P(\bar{\mathbb{Q}})} \\ &\leq \sum_{x,y,z} P(\mathbb{Q}\mathbb{K}) \log \frac{P(\mathbb{Q}|\mathbb{K})}{P(\bar{\mathbb{Q}}|\mathbb{K})}. \end{aligned} \tag{16}$$

This, along with (15)and (16), gives that

$$\begin{aligned} D(\mathbb{Q}\|\bar{\mathbb{Q}}) &\leq \sum_{x,y,z} P(\mathbb{Q}\mathbb{K}) \log \frac{P(\mathbb{Q}|\mathbb{K})}{P(\bar{\mathbb{Q}}|\mathbb{K})} \\ &= \sum_{x,y} P(\mathbb{K})P(\mathbb{Q}) \log P(\mathbb{Q}) \\ &\quad - \sum_{x,y,z} P(\mathbb{Q})P(\mathbb{K}) \log P(\bar{\mathbb{Q}}|\mathbb{K}) \\ &= -\sum_{x} P(\mathbb{K})\mathcal{H}(\mathbb{Q}) \\ &\quad - \sum_{x,z} P(\mathbb{K}) \log P(\bar{\mathbb{Q}}|\mathbb{K}) \\ &= -\mathcal{H}(\mathbb{Q}) \\ &\quad - \sum_{x,z} \frac{P(\mathbb{K})}{P(\bar{\mathbb{Q}})} P(\bar{\mathbb{Q}}) \log P(\bar{\mathbb{Q}}|\mathbb{K}). \end{aligned} \tag{17}$$

From theorem 1, when we get the target state, it satisfies

$$P(\mathbb{Q}|\bar{\mathbb{Q}}) = 1 \text{ and } P(\mathbb{K}|\bar{\mathbb{Q}}) = P(\mathbb{K}).$$

Hence, when $P(\mathbb{K}|\bar{\mathbb{Q}}) \to P(\mathbb{K})$,

$$P(\bar{\mathbb{Q}}|\mathbb{K}) = \frac{P(\bar{\mathbb{Q}}, \mathbb{K})}{P(\mathbb{K})} = \frac{P(\mathbb{K}|\bar{\mathbb{Q}})P(\bar{\mathbb{Q}})}{P(\mathbb{K})} \to P(\bar{\mathbb{Q}}).$$

Therefore, we get

$$-\sum_{x,z} P(\mathbb{K}) \log P(\bar{\mathbb{Q}}|\mathbb{K}) \to \mathcal{H}(\bar{\mathbb{Q}}).$$

This, together with (17), gives that when $P(\mathbb{K}|\bar{\mathbb{Q}}) \to P(\mathbb{K})$,

$$\begin{aligned} &D(\mathbb{Q}\|\bar{\mathbb{Q}}) \leq \mathcal{H}(\bar{\mathbb{Q}}) - \mathcal{H}(\mathbb{Q}), \\ &\text{as } \ P(\mathbb{K}|\bar{\mathbb{Q}}) \to P(\mathbb{K}). \end{aligned}$$

In addition,

$$\begin{aligned} &D(\mathbb{Q}\|\bar{\mathbb{Q}}) \leq \mathcal{H}(\bar{\mathbb{Q}}) - \mathcal{H}(\mathbb{Q}) = 0, \\ &\text{as } \ P(\mathbb{K}|\bar{\mathbb{Q}}) \to P(\mathbb{K}) \text{ and } P(\mathbb{Q}|\bar{\mathbb{Q}}) \to 1. \end{aligned} \tag{18}$$

On the other hand, by the non-negativity of KL divergence, we know that

$$D(\mathbb{Q}|\bar{\mathbb{Q}}) \geq 0.$$

From the Squeeze Theorem, it is obvious that,

$$\begin{aligned} D(\mathbb{Q}\|\bar{\mathbb{Q}}) = 0 \ \text{ as } \ P(\mathbb{K}|\bar{\mathbb{Q}}) \to P(\mathbb{K}) \\ \text{and } P(\mathbb{Q}|\bar{\mathbb{Q}}) \to 1. \end{aligned} \tag{19}$$

Further, the Pinsker's inequality (4) implies

$$D(\mathbb{Q}\|\bar{\mathbb{Q}}) \geq \frac{\log e}{2} \|P(\mathbb{Q}) - P(\bar{\mathbb{Q}})\|^2.$$

Then,

$$0 \leq \|P(\mathbb{Q}) - P(\bar{\mathbb{Q}})\| \leq \sqrt{\frac{2}{\log e} D(\mathbb{Q}\|\bar{\mathbb{Q}})}.$$

Similar to (19), we easily obtain

$$\begin{aligned} &\|P(\mathbb{Q}) - P(\bar{\mathbb{Q}})\| = 0 \\ &\text{as } \ P(\mathbb{K}|\bar{\mathbb{Q}}) \to P(\mathbb{K}) \\ &\text{and } P(\mathbb{Q}|\bar{\mathbb{Q}}) \to 1. \end{aligned} \tag{20}$$

In summary, the distribution captured by MISO is consistent with the prior distribution of the minority class. □