Galaxy Morphology Classification Using a Semi-Supervised Learning Algorithm Based on Dynamic Threshold postprint

Authors:

Date: 2023-09-20T00:00:00+00:00

Abstract

Machine learning has emerged as a pivotal technique for galaxy morphology classification, driven by the rapid growth of astronomical data. However, conventional supervised learning incurs substantial annotation costs, as it requires large quantities of labeled data to achieve satisfactory performance. FixMatch, a prominent semi-supervised learning algorithm, has become an essential tool for leveraging vast amounts of unlabeled data. Nevertheless, its performance deteriorates substantially when confronted with large-scale, imbalanced datasets, as FixMatch employs a static threshold for pseudo-label filtering. Therefore, this study proposes a Dynamic Threshold Alignment (DTA) algorithm built upon the FixMatch framework. First, the reliable pseudo-label ratio for the majority class is determined, and the corresponding ratios for the remaining classes are approximated accordingly. Second, the threshold for pseudo-label selection is dynamically computed based on the estimated reliable pseudo-label ratio for each category. This dynamic threshold mechanism mitigates accuracy bias across categories and enhances learning for minority classes. Experimental results demonstrate that for galaxy morphology classification tasks, the proposed algorithm significantly outperforms supervised learning. With only 100 labeled samples, the accuracy and F1-score are improved by 12.8% and 12.6%, respectively. Compared to popular semi-supervised algorithms such as FixMatch and MixMatch, the proposed method achieves superior classification performance while substantially reducing accuracy bias across categories. When the number of labeled data is 1000, the accuracy for the cigar-shaped smooth galaxy class with the fewest samples is improved by 25.87% compared to FixMatch.

Full Text

Abstract

Machine learning has become a crucial technique for classifying galaxy morphology due to the meteoric growth of galactic data. Unfortunately, traditional supervised learning incurs significant costs since it requires large amounts of labeled data to be effective. FixMatch, a semi-supervised learning algorithm, serves as a promising method for leveraging vast quantities of unlabeled data. Nevertheless, its performance degrades significantly when dealing with large, imbalanced datasets because FixMatch uses a fixed threshold to filter pseudo-labels. Therefore, this study proposes a dynamic threshold alignment (DTA) algorithm based on the FixMatch model. First, the reliable pseudo-label ratio for the class with the highest sample count is determined, and the reliable pseudo-label ratios for the remaining classes are approximated accordingly. Second, the threshold for selecting pseudo-labels is dynamically calculated for each category based on the predicted reliable pseudo-label ratio. By employing this dynamic threshold, the accuracy bias across categories is reduced and the learning of minority classes is improved. Experimental results show that in galaxy morphology classification tasks, the proposed algorithm significantly improves performance compared with supervised learning. When the amount of labeled data is 100, the accuracy and F1-score are improved by 12.8% and 12.6%, respectively. Compared with popular semi-supervised algorithms such as FixMatch and MixMatch, the proposed algorithm achieves better classification performance and greatly reduces accuracy bias across categories. When the labeled data size is 1000, the accuracy for cigar-shaped smooth galaxies—the class with the fewest samples—is improved by 37.94% compared to FixMatch.

Key words: methods: data analysis — techniques: image processing — galaxies: general — galaxies: structure

1 Introduction

Understanding galaxy morphology is essential for investigating galaxy evolution (Barchi et al. 2020). Galaxy morphology is closely related to the formation process of galaxies (Holwerda 2021). By studying the morphological features of galaxies, we can explore galaxy evolution, the distribution of dark matter, and the measurement of cosmological parameters, providing valuable information for our understanding of the cosmos (Wijesinghe et al. 2010; Salucci 2019; Parry et al. 2009). For example, spiral arm characteristics affect how giant molecular clouds (GMCs) form within spiral arms and how their mass functions (Bekki 2021).

Currently, there are many galaxy morphology classification schemes, including a visual classification system based on the visual characteristics of galaxies (Kartaltepe et al. 2015), a model-based classification system based on the brightness profiles of galaxies (Peng et al. 2002), and a non-model-based classification sys-

tem based on structural parameters of galaxy morphology (Lotz et al. 2004), among others. A well-known visual classification scheme is the Hubble sequence, which divides galaxies into three broad classes based on their visual features: elliptical galaxies, spiral galaxies, and lenticular galaxies (Hubble 1979). These broad classifications are further refined to achieve more detailed galaxy morphology classification, leading to additional categories such as irregular galaxies (Gallagher & Hunter 1984). Inspired by the Hubble sequence, the Galaxy Zoo decision tree was designed to classify galaxy morphology more comprehensively (Willett et al. 2013).

The classification of galaxies initially relied on visual assessment (De Vaucouleurs 1959, 1964). However, the volume of galaxy data has grown tremendously due to ongoing sky surveys, including the Sloan Digital Sky Survey (SDSS, York et al. 2000), the Hyper Suprime-Cam survey (HSC, Miyazaki et al. 2012), the Dark Energy Survey (DES, Collaboration et al. 2005), the Euclid Space Telescope (EST, Laureijs et al. 2011), and the Vera Rubin Observatory Legacy Survey of Space and Time (LSST, Ivezić et al. 2019). For example, the LSST can generate 36TB of data per night, totaling 500PB over its lifetime (Farias et al. 2020). Faced with such large volumes of data, completing visual classification of galaxies is challenging even when utilizing citizen science projects like Galaxy Zoo (Willett et al. 2013). Consequently, applying machine learning to classify galaxy morphology has become an optimal choice (Reza 2021). For example, Gupta et al. (2022) proposed an improved version of ResNet for galaxy classification. Li et al. (2023) designed a multi-scale convolutional neural network to extract multi-scale features from galaxy images, resulting in improved classification accuracy. Fang et al. (2023) introduced adaptive polar coordinate transformation to ensure consistent classification results for the same galaxy image. Various machine learning methods have contributed to this field, including those by Dunn et al. (2023), Wu et al. (2022), Ghosh et al. (2022), Zhang et al. (2022), Wei et al. (2022), and Hui et al. (2022). Among these, traditional supervised machine learning necessitates substantial amounts of labeled data for galaxy morphology classification (Barchi et al. 2020; Zhu et al. 2019), and manual data labeling is time-consuming and labor-intensive, increasing learning costs. Therefore, the use of semi-supervised approaches to fully exploit unlabeled data and improve classification model performance has emerged as an important research direction in galaxy morphology classification.

Currently, an increasing number of semi-supervised algorithms are being applied to astronomical data analysis. For instance, Ma et al. (2019) built an autoencoder based on the VGG-16 network that was first trained on large amounts of unlabeled data to learn how to extract galactic features, and then fine-tuned on a small amount of labeled data for radio galaxy morphological classification. Soroka et al. (2021) proposed a semi-supervised approach based on active learning and adversarial autoencoder models to address galaxy morphology classification. Slijepcevic et al. (2022) conducted semi-supervised research based on the radio galaxy classification network of Tang et al. (2019) utilizing transfer

learning as the baseline, demonstrating the precision and robustness of semi-supervised learning in radio galaxy classification. Ćiprijanović et al. (2022) created the DeepAstroUDA method, a general semi-supervised domain adaptation technique for astronomical applications that can identify non-overlapping classes in two separate galaxy datasets and even discover and cluster unidentified classes.

Semi-supervised learning (SSL) enhances learning performance by incorporating unlabeled data based on small-sample supervised learning (Berthelot et al. 2019). Today, deep semi-supervised learning (DSSL), which combines SSL and deep learning, has emerged as the most effective method for semi-supervised learning (Yang et al. 2022). Deep semi-supervised learning schemes can be categorized into three groups: consistency regularization-based semi-supervised learning, pseudo-labeling based semi-supervised learning, and semi-supervised deep learning techniques combining consistency regularization with pseudo-labels. The pseudo-label is regarded as the prediction label of unlabeled data by the model trained using trustworthy labeled data, and pseudo-labels with high probability participate in the model's training in the same way as labeled data (Lee et al. 2013). Semi-supervised deep learning techniques include MixMatch, ReMixMatch, and FixMatch, which combine consistency regularization and pseudo-labels and have become the most popular solutions (Berthelot et al. 2019; Sohn et al. 2020). Among these algorithms, FixMatch simplifies the application of pseudo-labels and unsupervised loss and has been shown to achieve the best performance on basic test datasets.

Even though the FixMatch model performs optimally, this is only possible with balanced and sufficient data quantities for each category. However, training data in deep learning applications is typically imbalanced, especially in the domain of astronomical data. For instance, the GZ2 dataset cited in this article contains only a small number of cigar-shaped galaxies. When confronted with imbalanced datasets, the model tends to learn more features of majority classes and fewer features of minority classes, resulting in accuracy bias where the majority class accuracy is higher and the minority class accuracy is lower. This problem is primarily caused by FixMatch's predetermined high threshold for semi-supervised learning, which ignores the learning progress of various classes. Consequently, models such as FlexMatch (Zhang et al. 2021), Adsh (Guo & Li 2022), and Dash (Xu et al. 2021), which are based on the FixMatch model, introduce dynamic thresholds that change with the learning status.

For example, FlexMatch proposes the idea of curricular pseudo-labels, a curriculum learning approach to leverage unlabeled data according to the model's learning status, where the dynamic threshold is a nonlinear mapping between the number of pseudo-labels for each class whose confidence exceeds the threshold and the current threshold. To improve learning for minority classes, Adsh dynamically adjusts thresholds by determining the pseudo-label filtering ratio for each class. Meanwhile, DARP (Kim et al. 2020), ABC (Lee et al. 2021), CReST (Wei et al. 2021), and others optimize the issue of data imbalance in

semi-supervised learning from the perspective of adjusting class distributions. Despite various semi-supervised studies, little attention has been paid to the issue of imbalanced data distribution in astronomical data, which can lead to accuracy biases in semi-supervised tasks across different categories.

Therefore, this paper proposes a semi-supervised method based on Dynamic Threshold Alignment (DTA) to address the issue of data imbalance in semi-supervised classification of galaxies. By establishing a class-specific threshold that changes dynamically with the learning state of each class, the DTA method improves upon the fixed high threshold in the FixMatch algorithm. This ensures that minority classes receive a greater number of unlabeled learning samples during training, minimizing accuracy biases in classification tasks. We carried out experiments utilizing galaxy images from the Galaxy Zoo Data Challenge Project on Kaggle based on the Galaxy Zoo 2 project (Willett et al. 2013) to measure these improvements. We compared the experimental results of the FixMatch algorithm, several well-known semi-supervised algorithms, and the DTA algorithm under various data quantities. The DTA algorithm performed better in most situations.

The structure of this paper is as follows. Section 2 describes the methodology, including evaluation metrics and the design of the DTA algorithm. Section 3 presents the experiments, introducing the experimental datasets, platform, data augmentation, baseline network, and comparison techniques. Results and discussion are presented in Section 4. Section 5 concludes the paper with a summary.

2 Methodology

The DTA algorithm improves upon the fixed high threshold used in FixMatch by setting independent dynamic thresholds for each galaxy category. This avoids the issue of losing correct pseudo-labels that can occur when using a fixed high threshold for all classes in FixMatch. By using dynamic thresholds, DTA enhances model robustness, reduces accuracy bias, and introduces more accurate pseudo-labels during the training process.

2.1.1 Fixed Threshold in FixMatch

The FixMatch semi-supervised learning technique employs a fixed threshold to filter reliable pseudo-labels. During training, pseudo-labels and consistency regularization principles are used. For labeled data, FixMatch trains a supervised model using cross-entropy loss and weak augmentation. The generated supervised model is then further trained on unlabeled data, with the unlabeled data being subjected to weak augmentation, strong augmentation, and cross-entropy loss, respectively [Figure 1: see original paper]. According to the consistency regularization principle, the same unlabeled data should yield the same classification results after both weak and strong augmentations. By minimizing cross-entropy loss, FixMatch brings the strong augmentation prediction results

closer to the pseudo-labels, which are generated based on the weak augmentation prediction results of unlabeled data.

In FixMatch, there are two types of loss functions: supervised loss for labeled data and unsupervised loss for unlabeled data. Suppose that FixMatch employs labeled data $\mathcal{X} = \{(x_b^l, y_b^l) : b \in (1, ..., B)\}$ with batch size B and unlabeled data $\mathcal{U} = \{x_b^u : b \in (1, ..., \mu B)\}$ with quantity μB, where μ is the proportion of unlabeled to labeled data. The loss function of FixMatch is defined as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u$$

where λ_u is a constant scalar hyperparameter denoting the importance of unsupervised loss; $\mathcal{L}_s$ indicates supervised loss; and $\mathcal{L}_u$ indicates unsupervised loss. The supervised loss $\mathcal{L}_s$ is the standard cross-entropy loss of weakly augmented labeled data compared to the true label, calculated as:

$$\mathcal{L}_s = \frac{1}{B} \sum_{b=1}^{B} H\left(y_b^l, f(y|\alpha(x_b^l); \theta)\right)$$

where $\alpha(\cdot)$ represents weak data augmentation and $f(y|\alpha(x_b^l); \theta) \in [0, 1]^k$ is the prediction probability of weakly augmented labeled data $\alpha(x_b^l)$ by the model with parameter θ. $H(\cdot, \cdot)$ is the cross-entropy function.

The unsupervised loss $\mathcal{L}_u$ for unlabeled data with strong augmentation is a standard cross-entropy loss between the pseudo-label $\widehat{y}_b^u$ and the predicted result y_b^u calculated by $f(y|A(x_b^u); \theta)$. The equations are as follows:

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{I}(\max(q_b) \geq \tau) H(\widehat{y}_b^u, y_b^u = f(y|A(x_b^u); \theta))$$

$$q_b = f(y|\alpha(x_b^u); \theta)$$

$$\widehat{y}_b^u = \text{argmax}(q_b)$$

where $\mathbb{I}(\cdot)$ is a filter function to ensure the reliability of pseudo-labels; τ stands for the threshold defined by FixMatch; q_b represents the prediction probability of model f with parameter θ; $A(x_b^u)$ and $\alpha(x_b^u)$ represent strong and weak augmentation for unlabeled data, respectively; $\widehat{y}_b^u$ is the unlabeled data's pseudo-label in one-hot probability distribution form produced by applying the argmax function to the probability prediction value q_b. Based on the principle of consistency regularization, the FixMatch algorithm obtains the unsupervised loss of unlabeled data using cross-entropy loss with the corresponding pseudo-label.

In the FixMatch algorithm, a fixed high threshold $\tau = 0.95$ is configured to ensure the reliability of pseudo-labels by screening pseudo-labels with high prediction confidence. However, the high threshold limits the number of pseudo-labels while maintaining their validity. Especially in the early training stages, excessively high thresholds lead to the loss of correct pseudo-labels in minority classes, further increasing the training gap between minority and majority classes, which is detrimental to model robustness. Therefore, a new dynamic threshold semi-supervised approach is needed to minimize the loss of accurate pseudo-labels without relying on a predetermined high threshold during training.

2.1.2 Dynamic Threshold Alignment Algorithm

The main premise of the dynamic threshold alignment technique is to consider the influence of the number of labeled data in each class on the learning effect while assuming uniform distribution of different classes within a batch. Consequently, by examining the percentage in the majority class, we can infer the proportion of reliable pseudo-labels in other classes. The algorithm dynamically determines the threshold for filtering pseudo-labels in each category based on these inferred proportions, addressing the shortcoming of using a fixed threshold in the FixMatch algorithm.

The practical flow of the algorithm is displayed in [Figure 2: see original paper]. First, the predicted results of unlabeled data are grouped by class, and the confidence of the predicted class is stored in an array and sorted in descending order. Then, based on the fixed high threshold of the majority class, the reliable pseudo-label ratio of the majority class is determined, and the reliable pseudo-label ratios of other classes are calculated based on the class distribution of labeled data. Finally, based on the reliable pseudo-label ratios of each class, reliable pseudo-labels are assigned from high to low confidence in the sorted prediction arrays. The confidence corresponding to the partition position becomes the new threshold.

(1) Reliable Pseudo-Label Ratio Calculation

The DTA approach first establishes a predefined high threshold τ_0 for the majority class, ensuring reliable pseudo-label screening. Based on this, the ratio of pseudo-labels with confidence higher than the threshold in unlabeled data predicted as the majority class by the model can be calculated—i.e., the reliable pseudo-label ratio of the majority class—as shown in:

$$\rho = \frac{\text{length}(A_0^{\geq \tau_0})}{\text{length}(A_0)} = \frac{\sum_{b=1}^{M} \mathbb{I}(\text{argmax}(f(y|\alpha(x_b^u);\theta)) = 0) \cdot \mathbb{I}(f(y=0|\alpha(x_b^u);\theta) \geq \tau_0)}{\sum_{b=1}^{M} \mathbb{I}(\text{argmax}(f(y|\alpha(x_b^u);\theta)) = 0)}$$

where ρ is the pseudo-label ratio of the majority class; M is the total number of unlabeled data; $\sum_{b=1}^{M} \mathbb{I}(f(y = 0|\alpha(x_b^u);\theta) \geq \tau_0)$ is the number of pseudo-

labels in unlabeled data predicted as the majority class with confidence higher than the threshold; A_0 stores the confidence of unlabeled data predicted as the majority class, with confidences arranged in descending order; and length(A_0) is the number of unlabeled data predicted as the majority class.

The reliable pseudo-label ratios of each class can be computed using the ratio of each class's count in labeled data to that of the majority class, along with the reliable pseudo-label ratio of the majority class, as shown in:

$$\rho_i = \rho \times \frac{N[i]}{N[0]}$$

where ρ_i is the reliable pseudo-label ratio of class i; ρ is obtained from the previous equation as the reliable pseudo-label ratio of the majority class; $N[i]$ is the number of samples in class i; and $N[0]$ is the number of samples in the majority class in the labeled data.

(2) Dynamic Threshold Calculation

Using the reliable pseudo-label ratios of each class obtained from the equation above and the confidence of the model's prediction on unlabeled data, the new threshold for each class can be calculated using:

$$\text{new-}\tau_c = A_c[\text{length}(A_c) \times \rho_i]$$

$$\text{length}(A_c) = \sum_{b=1}^{M} \mathbb{I}(\text{argmax}(f(y|\alpha(x_b^u); \theta)) = c)$$

where A_c is an array storing the confidence of unlabeled data predicted as class c, sorted in descending order, and length(A_c) is the number of unlabeled data predicted as class c.

The DTA algorithm uses this equation to determine the dynamic threshold new-τ_c for each class by establishing the pseudo-label screening ratio for each class. When the model has high confidence in the pseudo-labels of a minority class and the dynamic threshold new-τ_c is higher than the majority class threshold τ_0, new-τ_c will be set to τ_0 to introduce more correct pseudo-labels when the model is in a better learning state.

The DTA algorithm can select trusted pseudo-labels with relatively low confidence but high intra-class confidence by applying dynamic and independent thresholds for each class, minimizing the learning bias caused by imbalanced data during training.

2.2 Framework for Semi-Supervised Classification Using DTA Algorithm

The DTA technique is used in this semi-supervised training procedure to create dynamic thresholds for selecting trustworthy pseudo-labels for unlabeled data. The framework for semi-supervised training is shown in [Figure 3: see original paper]. In the early phases of model training, weak data augmentation is used to create an initial supervised model. At this point, the supervised loss is the sole component of the total loss because the DTA algorithm focuses on training the supervised model. When the labeled data reach a good initialization state—i.e., when the supervised loss is less than the appropriate threshold—the training of unlabeled data is introduced and pseudo-labels are generated for unlabeled data based on the initial model.

The DTA algorithm's pseudo-label screening must meet two requirements: first, the model prediction confidence must be higher than the threshold; second, the model's predicted probability for the corresponding unlabeled data must have low information entropy. Information entropy measures uncertainty: uncertainty decreases with increasing information entropy and increases with decreasing information entropy. When analyzing pseudo-labels using information entropy, lower information entropy indicates higher certainty of the model on the pseudo-label. The DTA algorithm adds the information entropy restriction to pseudo-label screening to boost label certainty. When training additionally includes unlabeled data, the total loss comprises both supervised loss and unsupervised loss, computed as in Equation (1). The DTA algorithm's unsupervised loss computation is:

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{I}\left(\max(q_b) \geq \hat{\tau}_{y_b^u} \text{ and } E_b^u < \tau_{\text{info}}\right) H\left(\hat{y}_b^u, f(y|A(x_b^u); \theta)\right)$$

where $\max(q_b)$ is the maximum confidence of the pseudo-label; $\hat{\tau}_{y_b^u}$ is the confidence threshold of the class corresponding to the pseudo-label $\hat{y}_b^u$; E_b^u is the information entropy of the model's prediction probability for the pseudo-label; and τ_{info} is the information entropy threshold.

2.3 Evaluation

Equations (13) to (16) outline the procedure for calculating evaluation metrics for binary classification tasks, which include accuracy, precision, recall, and F1-score. In these equations, TP represents true positive, FP represents false positive, TN represents true negative, and FN represents false negative.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

For the multi-classification task of galaxy morphologies, accuracy is the ratio of correctly predicted samples to the total number of samples, measuring the overall prediction accuracy. Precision, recall, and F1-score are calculated by taking the unweighted average of metrics for each class, known as macro precision, macro recall, and macro F1. The calculation equations are:

$$\text{macro precision} = \frac{1}{C} \sum_{i=1}^{C} \text{precision}_i$$

$$\text{macro recall} = \frac{1}{C} \sum_{i=1}^{C} \text{recall}_i$$

$$\text{macro F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i$$

where C represents the number of galaxy classes.

3 Experiments

3.1 Data Preparation

The data used in this study is derived from Galaxy Zoo 2 (GZ2), publicly available through the Galaxy Zoo Data Challenge Project on Kaggle. The dataset contains 61,578 galaxy images from SDSS DR7 (Data Release 7) and provides 37 parameters describing galaxy morphology. These parameter values range from 0 to 1 and represent the probability distribution of galaxy morphology across 11 classification tasks in the GZ2 decision tree (Willett et al. 2013). Higher values indicate stronger agreement among volunteer classifiers regarding a given galaxy's features, suggesting more reliable results.

To simplify the classification task, Zhu et al. (2019) screened five types of galaxies based on sample cleaning and selection criteria from Galaxy Zoo: completely round smooth, in-between smooth (between completely round and cigar-shaped), cigar-shaped smooth, edge-on, and spiral galaxies. Examples for each category are shown in [Figure 4: see original paper]. Following the sample

cleaning and selection criteria outlined by Zhu et al. (2019), we filtered these five types of galaxies to select reliable manual labels. The specific selection criteria are shown in . The selected dataset consists of 28,793 clean galaxy image samples, each with dimensions of $424 \times 424 \times 3$ pixels.

Within each category, the screened clean samples were split into training and testing sets in a 9:1 ratio. To evaluate the performance of the DTA method with varying labeled data sizes, six unique labeled datasets were constructed as presented in .

3.2 Data Augmentation

During semi-supervised training, weak and strong data augmentation were applied to unlabeled data, whereas only weak data augmentation was applied to labeled data.

3.2.1 Weak Data Augmentation In this experiment, galaxy images were subjected to various weak data augmentations, as depicted in [Figure 5: see original paper], including rotation, cropping, flipping, altering image properties, scaling, and translation. First, the image was randomly rotated from 0° to 360° and randomly flipped vertically and horizontally with 50% probability. Second, to extract galaxy morphology data from the image's center and remove extraneous background information, the image was arbitrarily center-cropped to size $s \times s \times 3$ with jittered size, where $s \in [160, 240]$. Third, the image's brightness, contrast, saturation, and hue were randomly altered with an offset range of 0 to 0.2. Finally, the image was translated horizontally or vertically by 0 to 2 pixels and resized to $98 \times 98 \times 3$ pixels. Simple center-cropping and scaling were applied to galaxy images in the validation set to meet model training requirements.

3.2.2 Strong Data Augmentation To prevent missing important morphological features in galaxy images, we eliminated the random image cropping procedure from the FixMatch algorithm's strong data augmentation. Similar to weak augmentation, strong augmentation involves larger adjustments to galaxy images. Images are flipped and rotated initially, then subjected to larger-scale jittering for center cropping, resulting in a randomly selected $s \times s \times 3$ size, where $s \in [160, 280]$. The third stage involves randomly adjusting hue, saturation, contrast, and brightness with offsets ranging from 0 to 0.4. Images are finally resized to $98 \times 98 \times 3$ pixels and translated 0 to 6 pixels horizontally or vertically.

3.3 Implementation Details

The semi-supervised galaxy classification based on the DTA algorithm was implemented using Python 3.8.5 and PyTorch 1.7.1. Experiments were conducted on a computer with 16 GB of RAM and 16 GB of VRAM, using Conda for GPU

acceleration. To validate the effectiveness of the DTA algorithm, numerous comparative experiments were carried out, including three types: semi-supervised learning, imbalanced semi-supervised learning, and supervised learning. Relevant algorithms were selected for comparison: FixMatch, MixMatch, and ReMixMatch as semi-supervised algorithms, and Adsh, DARP, and FlexMatch as imbalanced semi-supervised algorithms.

The EfficientNet-G3 deep neural network created by Wu et al. (2022) served as the foundational network. It is a lightweight deep neural network with fewer parameters that is effective for galaxy morphology classification. The low parameter count of EfficientNet-G3 prevents model overfitting in semi-supervised learning with limited labeled data.

EfficientNet-G3 was trained with a batch size of 16 for 50,000 iterations as the baseline network for all experiments. The ratio of unlabeled to labeled data during training was 7:1. The coefficient λ_u for unsupervised loss was set to 1. The threshold for supervised loss (τ_{loss}) was set to 0.2, and the threshold for information entropy (τ_{info}) was set to 0.4. Experiments used an SGD optimizer with a learning rate of 0.001 and an exponential moving average (EMA) approach with a decay rate of 0.999. The threshold τ_0 for the class with the largest number of samples was set to 0.95.

4 Results and Discussion

4.1 Results of DTA Algorithm and Baseline Network

EfficientNet-G3 served as our baseline network for both supervised and semi-supervised methods. Results comparing supervised learning and the DTA algorithm for galaxy classification are shown in . When there are 100 labeled data samples, the DTA algorithm outperforms supervised learning by 12.8% in accuracy and 12.6% in F1-score. Even with limited labels, semi-supervised learning achieves 91.8% accuracy. This demonstrates that the DTA method considerably enhances galaxy classification performance by introducing unlabeled data when labeled data is scarce. The performance of supervised classification gradually improves as the quantity of labeled samples increases, eventually producing results comparable to semi-supervised classification.

The trends of accuracy and F1-score with respect to labeled sample quantity are depicted in [Figure 6: see original paper] and [Figure 7: see original paper], respectively. For supervised learning, performance is significantly affected by the number of labels. Since semi-supervised learning can fully utilize unlabeled data, its performance is typically more consistent. There is only slight improvement when labeled data increases from 500 to 5000.

4.2 Comparison of DTA Algorithm with Other Semi-Supervised Algorithms

We selected six popular semi-supervised learning algorithms for comparison: FixMatch, MixMatch, ReMixMatch, Adsh, FlexMatch, and DARP. Comparative experimental results showing accuracy and F1-score for each model in galaxy classification tasks are presented in and . Visual comparisons are provided in [Figure 8: see original paper] and [Figure 9: see original paper] for more intuitive understanding. Overall, the DTA algorithm exceeds all other examined algorithms in accuracy and F1-score across most data scales.

When the labeled data size is 100, MixMatch achieves the best accuracy and F1-score. However, as galaxy data volume increases, MixMatch's F1-score drops sharply. [Figure 10: see original paper] presents MixMatch's recall rates for each galaxy category under different data scales. MixMatch adopts an aggressive data augmentation strategy, introducing more noise to the training set. While majority classes with abundant samples are less affected by noise, minority classes with fewer samples experience significant performance degradation. As galaxy data volume increases, the classification accuracy gap between minority and majority categories enlarges, and the recall rate for the least populated cigar-shaped smooth galaxies shows a downward trend, reaching 0 recall at scales of 1000, 2500, and 5000. Therefore, although MixMatch performs best at a data volume of 100, it does not generalize well to other scales for galaxy morphology classification. This problem stems from MixMatch's training strategy, which adds substantial noise through different random augmentations on the same unlabeled data. Since MixMatch and ReMixMatch rely heavily on data augmentation and fuse predictions from multiple random augmentations of the same image, we retained their original augmentation methods while using consistent augmentation for all other algorithms.

When the labeled data size is 250, DTA achieves the highest F1-score and its accuracy is second only to MixMatch. At a data size of 5000, DTA's F1-score is 1% lower than FlexMatch, but its accuracy reaches the highest at 95.6%. Across all data scales, DTA's accuracy and F1-score are higher than FixMatch, ReMixMatch, Adsh, and DARP. Meanwhile, DTA's and FlexMatch's accuracy steadily increase as labeled data size grows, closely tracking F1-score changes, with DTA outperforming FlexMatch at all scales. Consequently, DTA is a semi-supervised algorithm with good generalizability for galaxy morphology classification.

4.3 Visualization Analysis of DTA Algorithm and Other Algorithms

Since our algorithm is based on FixMatch, optimizing the fixed threshold to a dynamic threshold to address performance deterioration caused by data imbalance, our primary interest is investigating how dynamic thresholds affect classification improvement. [Figure 11: see original paper] shows confusion matrices on the validation set for DTA and other algorithms when the labeled data size is 1000. In confusion matrices, the diagonal where true and predicted

labels coincide represents the proportion of accurate predictions, while other values represent inaccurate predictions.

FixMatch's confusion matrix in [Figure 11: see original paper] illustrates classification accuracy bias in galaxy tasks, with cigar-shaped smooth galaxies performing poorly. Specifically, 82.76% of cigar-shaped smooth galaxies are misclassified as edge-on galaxies and 6.9% as in-between smooth galaxies. Edge-on galaxies are disk-shaped galaxies seen from the side, some with central bulges, while cigar-shaped smooth galaxies are a subtype of early-type galaxies that are smooth with small ellipticities. To avoid misclassification, we conducted cleaning and filtering to obtain clean samples and ensure correct manual labels during training. FixMatch's poor performance on cigar-shaped smooth galaxies is attributed to limited learning samples (only 1/6 of edge-on galaxies). Additionally, since both edge-on and cigar-shaped smooth galaxies have elliptical shapes, they may be confused if the model has insufficient training data.

To address limited learning samples for cigar-shaped smooth galaxies, the DTA algorithm dynamically adjusts the pseudo-label confidence threshold for each category during semi-supervised learning, as shown in [Figure 12: see original paper] (left). The threshold for cigar-shaped smooth galaxies is significantly lowered. Consequently, as shown in [Figure 12: see original paper] (right), more pseudo-labeled learning samples for cigar-shaped smooth galaxies are introduced during training, improving classification performance. The DTA algorithm significantly mitigates FixMatch's classification bias, increasing the accurate classification rate for cigar-shaped smooth galaxies by 37.94%, as seen in DTA's confusion matrix in [Figure 11: see original paper]. Improvements also occur for in-between smooth galaxies. This analysis demonstrates that DTA achieves more unbiased classification accuracy.

Comparing DTA's classification performance with other algorithms across galaxy categories, [Figure 11: see original paper] shows DTA outperformed all others in the minority class of cigar-shaped smooth galaxies with 48.28% classification accuracy. DTA also performed well on majority classes: completely round smooth galaxies (96.45% accuracy, higher than ReMixMatch, Adsh, and DARP), in-between smooth galaxies (93.93% accuracy, higher than all comparison algorithms), edge-on galaxies (97.18% accuracy, higher than supervised learning, MixMatch, Adsh, and DARP), and spiral galaxies (95.01% accuracy, higher than supervised learning and Adsh). Thus, DTA achieves good classification performance across all galaxy categories.

To explore the effect of dynamic thresholds on pseudo-label quantities, we created graphs showing threshold changes across iterations. [Figure 12: see original paper] (left) shows DTA's dynamic threshold adjustments. Different lines represent different galaxy types, with the vertical axis showing pseudo-label filtering thresholds. In early semi-supervised training stages, DTA lowers the threshold for cigar-shaped smooth galaxies, introducing more learning samples ([Figure 12: see original paper] right). This increases model performance. Analysis reveals that DTA dynamically adjusts thresholds based on sample distribution

across categories, effectively balancing training samples and enabling balanced accuracy across categories.

5 Conclusions

This study addresses semi-supervised learning for galaxy classification and proposes the DTA algorithm to handle data imbalance. Unlike FixMatch's constant threshold, DTA uses dynamic thresholds to improve learning of minority classes in semi-supervised training. Based on labeled data distribution, DTA calculates and aligns each galaxy category's classification performance with the most prevalent class, establishing each class's dynamic threshold through the total amount of added pseudo-labels. Experimental results demonstrate that DTA outperforms supervised learning and other well-known semi-supervised algorithms like FixMatch and MixMatch in enhancing classification performance and reducing accuracy bias across classes. Given the abundance of unlabeled data in large sky survey projects, the proposed DTA technique is highly important for galaxy morphology classification applications.

The DTA algorithm differs from other semi-supervised algorithms like DARP, ABC, and Adsh in that it does not need to consider unlabeled data distribution, preventing interference from incorrectly estimating unlabeled data distribution during training. DTA considers how labeled data distribution affects pseudo-label accuracy for unlabeled data, determining each class's dynamic threshold based on labeled data distribution and the percentage of trustworthy pseudo-labels in the most prevalent class.

Although DTA considerably improves classification performance for minority classes, their accuracy remains inferior to majority classes due to limited samples. Therefore, future work will focus on promoting learning for minority classes to achieve the same learning effect as the majority class, such as by introducing Generative Adversarial Networks (GAN).

Acknowledgements

This work was supported by China Manned Space Program through its Space Application System, the National Natural Science Foundation of China (Grant Nos. 11973022, U1811464), and the Natural Science Foundation of Guangdong Province (No. 2020A1515010710).

References

Barchi, P. H., de Carvalho, R., Rosa, R. R., et al. 2020, *Astronomy and Computing*, 30, 100334
Bekki, K. 2021, *Astronomy & Astrophysics*, 647, A120
Berthelot, D., Carlini, N., Goodfellow, I., et al. 2019, *Advances in Neural Information Processing Systems*
Ćiprijanović, A., Lewis, A., Pedro, K., et al. 2022, arXiv preprint arXiv:2211.00677

Collaboration, D. E. S., et al. 2005, arXiv preprint astro-ph/0510346
De Vaucouleurs, G. 1959, in *Astrophysik IV: Sternsysteme/Astrophysics IV: Stellar Systems* (Springer), 275
De Vaucouleurs, G. 1964, *Astronomical Journal*, Vol. 69, p. 737 (1964), 69, 737
Dunn, M. M., Ciprijanovic, A. M., Nord, B., & Mobasher, B. 2023, *Galaxy Morphology Classification Using Bayesian Neural Networks for LSST*, Tech. rep., Fermi National Accelerator Lab. (FNAL), Batavia, IL (United States)
Fang, G., Ba, S., Gu, Y., et al. 2023, *The Astronomical Journal*, 165, 35
Farias, H., Ortiz, D., Damke, G., Arancibia, M. J., & Solar, M. 2020, *Astronomy and Computing*, 33
Gallagher, J. S., & Hunter, D. A. 1984, *Annual Review of Astronomy and Astrophysics*, 22, 37
Ghosh, A., Urry, C. M., Rau, A., et al. 2022, *The Astrophysical Journal*, 935, 138
Guo, L.-Z., & Li, Y.-F. 2022, in *International Conference on Machine Learning*, PMLR, 8082
Gupta, R., Srijith, P., & Desai, S. 2022, *Astronomy and Computing*, 38, 100543
Holwerda, B. W. 2021, *Galaxy Morphology* (IOP Publishing)
Hubble, E. P. 1979, in *A Source Book in Astronomy and Astrophysics, 1900–1975* (Harvard University Press), 716
Hui, W., Jia, Z. R., Li, H., & Wang, Z. 2022, in *Journal of Physics: Conference Series*, IOP Publishing
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, *The Astrophysical Journal*, 873, 111
Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2015, *The Astrophysical Journal Supplement Series*, 221, 11
Kim, J., Hur, Y., Park, S., et al. 2020, *Advances in Neural Information Processing Systems*, 33, 14567
Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv preprint arXiv:1110.3193
Lee, D.-H., et al. 2013, in *Workshop on Challenges in Representation Learning*, ICML, Atlanta, 896
Lee, H., Shin, S., & Kim, H. 2021, *Advances in Neural Information Processing Systems*, 34, 7082
Li, G., Xu, T., Li, L., et al. 2023, *Monthly Notices of the Royal Astronomical Society*, 523, 488
Lotz, J. M., Primack, J., & Madau, P. 2004, *The Astronomical Journal*, 128, 163
Ma, Z., Zhu, J., Zhu, Y., & Xu, H. 2019, in *Data Mining and Big Data: 4th International Conference, DMBD 2019*, Chiang Mai, Thailand, July 26–30, 2019, Proceedings 4, Springer, 191
Miyazaki, S., Komiyama, Y., Nakaya, H., et al. 2012, in *Ground-based and Airborne Instrumentation for Astronomy IV*, Vol. 8446, SPIE, 327
Parry, O., Eke, V., & Frenk, C. 2009, *Monthly Notices of the Royal Astronomical Society*, 396, 1972
Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, *The Astronomical Journal*, 124, 266

Reza, M. 2021, *Astronomy and Computing*, 37, 100492
Salucci, P. 2019, *The Astronomy and Astrophysics Review*, 27, 1
Slijepcevic, I. V., Scaife, A. M., Walmsley, M., et al. 2022, *Monthly Notices of the Royal Astronomical Society*, 514, 2599
Sohn, K., Berthelot, D., Carlini, N., et al. 2020, *Advances in Neural Information Processing Systems*, 33
Soroka, A., Meshcheryakov, A., & Gerasimov, S. 2021, arXiv preprint arXiv:2105.02958
Tang, H., Scaife, A. M., & Leahy, J. 2019, *Monthly Notices of the Royal Astronomical Society*, 488, 3358
Wei, C., Sohn, K., Mellina, C., Yuille, A., & Yang, F. 2021, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10857
Wei, S., Li, Y., Lu, W., et al. 2022, *Publications of the Astronomical Society of the Pacific*, 134, 114508
Wijesinghe, D., Hopkins, A., Kelly, B., Welikala, N., & Connolly, A. 2010, *Monthly Notices of the Royal Astronomical Society*, 404, 2077
Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, *Monthly Notices of the Royal Astronomical Society*, 435, 2835
Wu, D., Zhang, J., Li, X., & Li, H. 2022, *Research in Astronomy and Astrophysics*, 22, 115011
Xu, Y., Shang, L., Ye, J., et al. 2021, in *International Conference on Machine Learning*, PMLR, 11525
Yang, X., Song, Z., King, I., & Xu, Z. 2022, *IEEE Transactions on Knowledge and Data Engineering*
York, D. G., Adelman, J., Anderson Jr, J. E., et al. 2000, *The Astronomical Journal*, 120, 1579
Zhang, B., Wang, Y., Hou, W., et al. 2021, *Advances in Neural Information Processing Systems*, 34, 18408
Zhang, Z., Zou, Z., Li, N., & Chen, Y. 2022, *Research in Astronomy and Astrophysics*, 22, 055002
Zhu, X.-P., Dai, J.-M., Bian, C.-J., et al. 2019, *Astrophysics and Space Science*, 364, 1

Note: Figure translations are in progress. See original paper for figures.