ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202312.00144

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

Authors: Jie Jiang, Jinqu Zhang, Xiangru Li, Hui Li and Ping Du

Date: 2023-12-15T00:00:00+00:00

Abstract

Machine learning has become a crucial technique for classifying the morphology of galaxies as a result of the meteoric development of galactic data. Unfortunately, traditional supervised learning has significant learning costs since it needs a lot of labeled data to be effective. FixMatch, a semi-supervised learning algorithm that serves as a good method, is now a key tool for using large amounts of unlabeled data. Nevertheless, the performance degrades significantly when dealing with large, imbalanced data sets since FixMatch relies on a fixed threshold to filter pseudo-labels. Therefore, this study proposes a dynamic threshold alignment algorithm based on the FixMatch model. First, the class with the highest amount has its reliable pseudo-label ratio determined, and the remaining classes' reliable pseudo-label ratios are approximated in accordance. Second, based on the predicted reliable pseudo-label ratio for each category, it dynamically calculates the threshold for choosing pseudo-labels. By employing this dynamic threshold, the accuracy bias of each category is decreased and the learning of classes with less samples is improved. Experimental results show that in galaxy morphology classification tasks, compared with supervised learning, the proposed algorithm significantly improves performance. 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 has better classification performance, greatly reducing the accuracy bias of each category. When the amount of labeled data is 1000, the accuracy of cigar-shaped smooth galaxies with the smallest sample is improved by 37.94% compared to FixMatch.

ChinaRxiv 中X

Full Text

Preamble

Research in Astronomy and Astrophysics, 23:115019 (14pp), 2023 November
© 2023. National Astronomical Observatories, CAS and IOP Publishing Ltd. Printed in China and the U.K.
https://doi.org/10.1088/1674-4527/acf610

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

Jie Jiang[1], Jinqu Zhang[1], Xiangru Li[1], Hui Li[1], and Ping Du[2]

[1] School of Computer Science, South China Normal University, Guangzhou 510631, China; zjq@scnu.edu.cn

[2] Guangdong Construction Vocational Technology Institute, Qingyuan 511500, China

Received 2023 May 2; revised 2023 August 17; accepted 2023 August 24; published 2023 October 11

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 because it requires large amounts of labeled data to be effective. FixMatch, a semi-supervised learning algorithm, has emerged as a key tool for leveraging vast quantities of unlabeled data. Nevertheless, its performance degrades substantially when dealing with large, imbalanced datasets since FixMatch relies on a fixed threshold to filter pseudo-labels. Therefore, this study proposes a dynamic threshold alignment algorithm based on the FixMatch model. First, the reliable pseudo-label ratio is determined for the class with the highest sample count, and the reliable pseudo-label ratios for the remaining classes are approximated accordingly. Second, the threshold for selecting pseudo-labels is dynamically calculated based on the predicted reliable pseudo-label ratio for each category. By employing this dynamic threshold, the accuracy bias across categories is reduced and the learning of minority classes is improved. Experimental results demonstrate 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 while greatly reducing accuracy bias across categories. When the amount of labeled data is 1000, the accuracy for cigar-shaped smooth galaxies—the class with the smallest sample size—is improved by 37.94% compared to FixMatch.

Key words: galaxies: photometry – techniques: image processing – techniques: photometric

1. Introduction

Investigating the evolution of galaxies requires an understanding of galaxy morphology [**?**]. Galaxy morphology is closely related to the formation process of galaxies [**?**]. By studying the morphological features of galaxies, we can explore the evolution of galaxies, the distribution of dark matter, and the measurement of cosmological parameters, providing valuable information for our understanding of the cosmos [**?**, **?**, **?**]. For example, spiral arm characteristics affect how giant molecular clouds form within spiral arms and how their mass is distributed [**?**].

Currently, there are many galaxy morphology classification schemes, including a visual classification system based on the visual characteristics of galaxies [**?**], a model-based classification system based on the brightness profiles of galaxies [**?**], a non-model-based classification system based on structural parameters of galaxy morphology [**?**], and 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 [**?**, **?**, **?**]. These broad classifications are further refined to achieve more detailed galaxy morphology classification, leading to the development of additional categories like irregular galaxies [**?**].

Inspired by the Hubble sequence, the Galaxy Zoo decision tree was designed to classify galaxy morphology in a more comprehensive way [**?**]. The classification of galaxies initially relied on visual assessment. However, the amount of galaxy data has grown tremendously due to ongoing sky surveys, including the Sloan Digital Sky Survey (SDSS; [**?**]), the Hyper Suprime-Cam (HSC) survey [**?**], the Dark Energy Survey [**?**], the Euclid Space Telescope (EST; [**?**]), and the Vera Rubin Observatory Legacy Survey of Space and Time (LSST; [**?**]). For example, the LSST can generate 36 TB of data per night, totaling 500 PB over its lifetime [**?**]. Faced with such a large volume of data, it is challenging to complete visual classification even utilizing citizen science projects like Galaxy Zoo [**?**]. Consequently, applying machine learning to classify galaxy morphology has become the best choice [**?**]. For example, [**?**] proposed an improved version of ResNet for galaxy classification. [**?**] designed a multi-scale convolutional neural network to extract multi-scale features from galaxy images, resulting in improved accuracy. [**?**] introduced adaptive polar coordinate transformation to ensure consistent classification results for the same galaxy image. Different machine learning methods have also contributed to this field, such as those by [**?**, **?**, **?**, **?**, **?**]. Among them, traditional supervised machine learning necessitates substantial amounts of labeled data for galaxy morphology classification [**?**, **?**], 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, more semi-supervised algorithms are being applied to astronomical data analysis. For instance, [**?**] 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. [**?**] suggested a semi-supervised approach based on active learning and adversarial autoencoder models for classifying galaxy morphologies. [**?**] conducted semi-supervised research based on the radio galaxy classification network of [**?**], utilizing transfer learning as the baseline and demonstrating the precision and robustness of semi-supervised learning in radio galaxy classification. [**?**] created the DeepAstroUDA method, a general semi-supervised domain adaptation technique for astronomical applications that can find non-overlapping classes in two separate galaxy datasets and even find and cluster unidentified classes.

Semi-supervised learning enhances performance by incorporating unlabeled data based on small-sample supervised learning [**?**]. Today, deep semi-supervised learning (DSSL), which combines SSL and deep learning, has emerged as the most effective method [**?**]. DSSL schemes can be categorized into three groups: consistency regularization-based SSL, pseudo-labeling-based SSL, and techniques combining both principles. A pseudo-label is regarded as a prediction label for unlabeled data by a model trained using trustworthy labeled data, and pseudo-labels with high confidence participate in model training similarly to labeled data [**?**]. Semi-supervised deep learning techniques include MixMatch, ReMixMatch, and FixMatch, which combine consistency regularization and pseudo-labels and have become the most popular solutions [**?**, **?**]. Among these algorithms, FixMatch simplifies the application of pseudo-labels and unsupervised loss and has been shown to achieve the best performance on standard benchmark datasets.

Even though FixMatch performs well, this is only possible with balanced and sufficient data quantities for each category. However, training data in deep learning applications are typically imbalanced, especially in astronomical data. For instance, the Galaxy Zoo 2 (GZ2) dataset used in this study contains only a small number of cigar-shaped galaxies. When confronted with imbalanced datasets, models tend to learn more features from majority classes and fewer features from minority classes, resulting in accuracy bias where majority class accuracy is higher and minority class accuracy is lower. This problem is primarily caused by FixMatch's fixed high threshold for SSL, which ignores the learning progress of different classes. Consequently, models like FlexMatch [**?**], Adsh [**?**], and Dash [**?**], which are based on FixMatch, introduce dynamic thresholds that change with learning status. For example, FlexMatch proposes curricular pseudo-labels, a curriculum learning approach that leverages unlabeled data according to the model's learning status, where the dynamic threshold represents

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 [**?**], ABC [**?**], CReST [**?**], and others optimize the issue of data imbalance in SSL by 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 data imbalance in semi-supervised galaxy classification. 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, thereby minimizing accuracy biases in classification tasks. We conducted experiments using galaxy images from the Galaxy Zoo Data Challenge Project on Kaggle based on the GZ2 project [**?**] to evaluate 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 experimental setup, introducing the 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 an independent dynamic threshold for each galaxy category. This avoids the issue of losing correct pseudo-labels that can occur when relying on a fixed high threshold for all classes in FixMatch. By utilizing a dynamic threshold, DTA enhances model robustness, reduces accuracy bias, and introduces more accurate pseudo-labels during training.

2.1. Dynamic Threshold Calculation

2.1.1. Fixed Threshold in FixMatch The FixMatch SSL technique employs a fixed threshold to filter reliable pseudo-labels during training, using both pseudo-labels and consistency regularization principles. 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 (Figure 1). According to the consistency

regularization principle, the same unlabeled data should yield identical 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 $\{(x_b, y_b)\}$ with a batch size of B and unlabeled data $\{u_b\}$ with a quantity of μB, where μ is the proportion of unlabeled to labeled data. The loss function of FixMatch is defined as follows:

$$\mathcal{L} = \mathcal{L}_s + \lambda_u \mathcal{L}_u \quad (1)$$

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

$$\mathcal{L}_s = \frac{1}{B} \sum_{b=1}^{B} H(y_b, p_b^w) \quad (2)$$

where $\alpha(\cdot)$ represents weak augmentation of labeled data; $p_b^w = f(\alpha(x_b); \theta)$ is the prediction probability of weakly augmented labeled data x_b by the model with parameter θ; and $H(\cdot, \cdot)$ 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 $\hat{y}_b$ and the predicted result q_b calculated by $f(\mathcal{A}(u_b); \theta)$. The equations are listed as follows:

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}(\max(q_b) \geq \tau) H(\hat{y}_b, q_b) \quad (3)$$

$$q_b = f(\mathcal{A}(u_b); \theta) \quad (4)$$

$$\hat{y}_b = \arg\max(q_b) \quad (5)$$

where $\mathbb{1}(\cdot)$ is a filter function to ensure the reliability of pseudo-labels; τ stands for the threshold defined by FixMatch; f represents the model with parameter θ; $\mathcal{A}$ and α signify strong and weak augmentation for unlabeled data, respectively; $\hat{y}_b$ is the unlabeled data's pseudo-label in one-hot probability distribution form, produced by applying the $\arg\max(\cdot)$ function to the probability prediction value q_b. Based on the principle of consistent 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 pseudo-label reliability by screening pseudo-labels with high prediction confidence. However, this high threshold limits the number of pseudo-labels while maintaining validity. Especially in the early training stages, excessively high thresholds lead to 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 must be implemented to minimize the loss of accurate pseudo-labels.

2.1.2. Dynamic Threshold Alignment Algorithm The main premise of the DTA technique is to consider the influence of the number of labeled data in each class on the learning effect while assuming a uniform distribution of different classes within a batch. Consequently, by examining the percentage in the majority class, we may 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 FixMatch.

The practical flow of the algorithm is displayed in Figure 2 [Figure 2: see original paper]. First, the predicted results of unlabeled data are grouped by class, and the confidence of each 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 can be calculated—that is, the reliable pseudo-label ratio of the majority class—as shown in:

$$\rho = \frac{\text{length}(\{q \mid \arg\max(q) = y_{\text{major}} \wedge \max(q) \geq \tau_0\})}{\text{length}(\{q \mid \arg\max(q) = y_{\text{major}}\})} \quad (8)$$

where ρ is the pseudo-label ratio of the majority class; M is the total number of unlabeled data; length$(\cdot)$ is the number of unlabeled data predicted as the

majority class; and A_0 stores the confidence values of unlabeled data predicted as the majority class, sorted in descending order.

The reliable pseudo-label ratios of each class can be computed using the ratio of labeled data counts relative to the majority class and the reliable pseudo-label ratio of the majority class:

$$\rho_i = \rho \times \frac{N[i]}{N[0]} \quad (9)$$

where ρ_i is the reliable pseudo-label ratio of class i; ρ is obtained from Equation (8) 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 Equation (9) and the model's prediction confidence on unlabeled data, the new threshold for each class can be calculated as:

$$\text{new-}\tau_c = A_c[\lfloor \rho_c \times \text{length}(A_c) \rfloor] \quad (10)$$

where A_c is an array storing the confidence of unlabeled data predicted as class c, sorted in descending order; $\text{length}(A_c)$ is the number of unlabeled data predicted as class c; and $\lfloor \cdot \rfloor$ denotes the floor function.

The DTA algorithm uses Equation (10) to determine the dynamic threshold new-τ_c for each class by establishing the pseudo-label screening ratio. 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.

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

2.2. Framework for Semi-supervised Classification Using DTA Algorithm

The DTA technique is employed 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 illustrated in Figure 3 [Figure 3: see original paper].

Weak data augmentation is used to create an initial supervised model in the early phases of training. At this point, the total loss consists only of supervised loss because the DTA algorithm focuses on training the supervised model. When the labeled data reach a good initialization state—specifically, when the supervised loss is less than threshold_{Ls}—the training of unlabeled data is introduced, and pseudo-labels are generated 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 predicted probability distribution of the corresponding unlabeled data must have low information entropy. Information entropy measures uncertainty; uncertainty decreases with lower information entropy. When analyzing pseudo-labels using information entropy, lower entropy indicates higher model certainty on the pseudo-label. The DTA algorithm adds the information entropy restriction to pseudo-label screening to boost label certainty.

When unlabeled data are included in training, the total loss comprises both supervised and unsupervised loss, computed as in Equation (1). The DTA algorithm's unsupervised loss calculation is:

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}(\max(q_b) \geq \tau_{c_b} \wedge \text{info\_entropy}(q_b) \leq \text{info\_}\tau) H(\hat{y}_b, q_b) \quad (11)$$

where τ_{c_b} is the confidence threshold of the class corresponding to the pseudo-label $\hat{y}_b$; info_entropy(q_b) is the information entropy of the model's prediction probability for the pseudo-label; and info_τ is the information entropy threshold.

2.3. Evaluation Metrics

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

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (14)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (15)$$

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (16)$$

For the multi-classification task of galaxy morphologies, accuracy is the ratio of correctly predicted samples to the total number of samples, measuring 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}:

$$\text{macro\_precision} = \frac{1}{C} \sum_{i=1}^{C} \text{precision}_i \quad (17)$$

$$\text{macro\_recall} = \frac{1}{C} \sum_{i=1}^{C} \text{recall}_i \quad (18)$$

$$\text{macro\_F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i \quad (19)$$

where C represents the number of galaxy classes.

3. Experiments

3.1. Data Preparation

The data used in this study are derived from GZ2, publicly available through the Galaxy Zoo Data Challenge Project on Kaggle. The dataset contains 61,578 galaxy images from SDSS Data Release 7 (DR7) 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 [**?**]. Higher values indicate stronger agreement among volunteer classifiers regarding a given galaxy's features, suggesting more reliable results.

To simplify the classification task, five galaxy types were selected by [**?**] based on Galaxy Zoo's sample cleaning and selection criteria: 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 depicted in Figure 4 [Figure 4: see original paper]. Following the sample cleaning and selection criteria outlined by [**?**], we filtered these five galaxy types to select reliable manual labels. The specific selection criteria are shown in Table 1 . The resulting dataset consists of 28,793 clean galaxy image samples, each with dimensions of 424 × 424 × 3 pixels.

Within each category, the clean samples were split into training and testing sets in a 9:1 ratio. To evaluate DTA performance with varying labeled data sizes, six unique labeled datasets were constructed as presented in Table 2 .

3.2. Data Augmentation

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

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

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

3.3. Implementation Details

Using Python 3.8.5 and PyTorch 1.7.1, the SSL galaxy classification based on DTA was implemented on a computer with 16 GB RAM and 16 GB VRAM, utilizing Conda for GPU acceleration. To validate DTA effectiveness, three types of comparative experiments were conducted: standard SSL, imbalanced SSL, and supervised learning. For comparison, we selected FixMatch, MixMatch, and ReMixMatch as semi-supervised algorithms, and Adsh, DARP, and FlexMatch as imbalanced semi-supervised algorithms.

The EfficientNet-G3 deep neural network [**?**] served as our baseline—a lightweight network with fewer parameters that is effective for galaxy morphology classification. Its low parameter count prevents overfitting in SSL with limited labeled data. EfficientNet-G3 was trained with batch size 16 for 50,000

iterations. The ratio of unlabeled to labeled data was 7:1. The unsupervised loss coefficient λ_u was set to 1. The supervised loss threshold loss_τ was 0.2, and the information entropy threshold info_τ was 0.4. We used SGD optimizer with learning rate 0.001 and exponential moving average (EMA) with decay rate 0.999. The threshold τ_0 for the majority class was set to 0.95.

4. Results and Discussion

4.1. Results of DTA Algorithm and Baseline Network

EfficientNet-G3 was our baseline for both supervised and semi-supervised methods. Table 3 compares supervised learning and DTA algorithm performance. With 100 labeled samples, DTA outperforms supervised learning by 12.8% in accuracy and 12.6% in F1-score, achieving 91.8% accuracy even with limited labels. This demonstrates that DTA considerably enhances galaxy classification performance by introducing unlabeled data when labeled data are scarce. Supervised classification performance gradually improves as labeled sample quantity increases, eventually producing results comparable to semi-supervised classification.

Figures 6 [Figure 6: see original paper] and 7 [Figure 7: see original paper] depict accuracy and F1-score trends relative to labeled sample quantity. Supervised learning performance is significantly affected by label count, while SSL performance remains relatively consistent due to its ability to fully utilize unlabeled data. Performance improvement is slight but not appreciable when labeled data increase from 500 to 5000.

4.2. Comparison of DTA Algorithm with Other Semi-supervised Algorithms

We compared DTA against six popular SSL algorithms: FixMatch, MixMatch, ReMixMatch, Adsh, FlexMatch, and DARP. Tables 4 and 5 show accuracy and F1-score for each model, with Figures 8 [Figure 8: see original paper] and 9 [Figure 9: see original paper] providing visual comparisons. Overall, DTA exceeds all other algorithms in accuracy and F1-score across most data scales.

With 100 labeled samples, MixMatch achieves the highest accuracy and F1-score, but its F1-score drops sharply as data volume increases. Figure 10 [Figure 10: see original paper] shows MixMatch's recall rates per galaxy category at different scales. MixMatch's aggressive augmentation strategy introduces significant noise to the training set. While majority classes with abundant samples are less affected, minority classes suffer substantial performance degradation. As galaxy data volume increases, the classification accuracy gap between minority and majority classes widens, with cigar-shaped smooth galaxies showing a downward recall trend, reaching 0% recall at scales of 1000, 2500, and 5000. Thus, although MixMatch performs best at data volume 100, it does not generalize well to other scales for galaxy morphology classification. This problem

stems from MixMatch's approach of applying different random augmentations to the same unlabeled data, adding substantial noise. Since MixMatch and ReMixMatch rely heavily on data augmentation and require fusing predictions from multiple random augmentations, we retained their original augmentation methods while using consistent augmentation for all other algorithms.

At 250 labeled samples, DTA achieves the highest F1-score and accuracy second only to MixMatch. At 5000 samples, DTA's F1-score is 1% lower than FlexMatch's, but its accuracy reaches the highest at 95.6%. Across all scales, DTA's accuracy and F1-score exceed those of FixMatch, ReMixMatch, Adsh, and DARP. FlexMatch's performance steadily increases with labeled data size, closely tracking F1-score changes, but DTA's accuracy outperforms FlexMatch at all scales. Consequently, DTA demonstrates good generalizability for galaxy morphology classification.

To investigate how dynamic thresholds affect classification improvement, Figure 11 [Figure 11: see original paper] shows confusion matrices on the validation set for DTA and other methods when labeled data size is 1000. The diagonal represents the proportion of accurate predictions, while off-diagonal elements show misclassification proportions. FixMatch's confusion matrix exhibits clear classification accuracy bias, with poor performance on cigar-shaped smooth galaxies—the minority class. 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 viewed 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. Despite sample cleaning and filtering to ensure correct manual labels, FixMatch performs poorly on cigar-shaped smooth galaxies due to limited learning samples (only 1/6 of edge-on galaxies) and visual similarity between edge-on and cigar-shaped smooth galaxies.

To address limited learning samples for cigar-shaped smooth galaxies, DTA dynamically adjusts the pseudo-label confidence threshold for each category during SSL, significantly lowering the threshold for cigar-shaped smooth galaxies (Figure 12 [Figure 12: see original paper], left). This introduces more pseudo-labeled learning samples for cigar-shaped smooth galaxies during training (Figure 12 [Figure 12: see original paper], right), improving classification performance. The DTA confusion matrix in Figure 11 shows this significantly increases the correct classification rate for cigar-shaped smooth galaxies by 37.94%, addressing FixMatch's biased classification issue. Improvements are also observed for in-between smooth galaxies. This demonstrates that DTA achieves more unbiased classification accuracy across all categories.

Comparing classification performance across galaxy categories, DTA outperforms all other algorithms, achieving 48.28% accuracy on the minority cigar-shaped smooth class. DTA also performs well on majority classes: 96.45% on completely round smooth galaxies (higher than ReMixMatch, Adsh, and DARP), 93.93% on in-between smooth galaxies (higher than all comparison algorithms), 97.18% on edge-on galaxies (higher than supervised learning, Mix-

Match, Adsh, and DARP), and 95.01% on spiral galaxies (higher than supervised learning and Adsh). Thus, DTA achieves good classification performance across all galaxy categories.

4.3. Visualization Analysis of DTA Algorithm and Other Algorithms

Since our algorithm optimizes FixMatch's fixed threshold to a dynamic threshold to address performance deterioration from data imbalance, we investigated how dynamic thresholds affect classification improvement. Figure 12 (left) shows dynamic threshold adjustments across training iterations. In early semi-supervised training stages, DTA lowers the threshold for cigar-shaped smooth galaxies, introducing more learning samples (Figure 12, right). This inclusion of additional training samples improves 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 SSL application for galaxy classification and proposes the DTA algorithm to handle data imbalance. DTA implements dynamic thresholds instead of FixMatch's constant threshold to improve minority class learning in semi-supervised training. Based on labeled data distribution, DTA calculates classification performance for each galaxy type and aligns it 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 classification performance while 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.

DTA 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 this 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 enhances classification performance for low-sample classes, their accuracy remains inferior to high-sample classes due to limited samples. Future work will focus on promoting learning for low-sample classes, such as by introducing Generative Adversarial Networks to achieve the same learning effect as the majority class.

Acknowledgments

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

ORCID iDs Jinqu Zhang https://orcid.org/0000-0001-6643-4053

References

Abbott, T., Aldering, G., & Annis, J. 2005, The Dark Energy Survey, arXiv: astro-ph/0510346

Barchi, P. H., de Carvalho, R., Rosa, R. R., et al. 2020, A&C, 30, 100334

Bekki, K. 2021, A&A, 647, A120

Berthelot, D., Carlini, N., Goodfellow, I., et al. 2019, Advances in Neural Information Processing Systems 32 (NeurIPS 2019) (Vancouver: IEEE), 32

Ćiprijanović, A., Lewis, A., Pedro, K., et al. 2022, arXiv:2211.00677

De Vaucouleurs, G. 1959, Classification and Morphology of External Galaxies, in Astrophysik iv: Sternsysteme/astrophysics iv: Stellar Systems, ed. S. Flügge (Berlin: Springer), 275

De Vaucouleurs, G. 1964, AJ, 69, 737

Dunn, M. M., Ciprijanovic, A. M., Nord, B., & Mobasher, B. 2023, Galaxy Morphology Classification Using Bayesian Neural Networks for LSST FERMILAB-POSTER-23-001-SCD, Fermi National Accelerator Lab. (FNAL), Batavia, IL

Fang, G., Ba, S., Gu, Y., et al. 2023, AJ, 165, 35

Farias, H., Ortiz, D., Damke, G., Arancibia, M. J., & Solar, M. 2020, A&C, 33

Gallagher, J. S., & Hunter, D. A. 1984, ARA&A, 22, 37

Ghosh, A., Urry, C. M., Rau, A., et al. 2022, ApJ, 935, 138

Guo, L., & Li, Y. 2022, PMLR, 162, 8082

Gupta, R., Srijith, P., & Desai, S. 2022, A&C, 38, 100543

Holwerda, B. W. 2021, Galaxy Morphology (Bristol: IOP Publishing)

Hou, W., Okumura, M., Shinozaki, T., et al. 2021, in Advances in Neural Information Processing Systems 34, ed. M. Ranzato et al., 34 (Online: NIPS), 18408

Hubble, E. P. 1979, in A Source Book in Astronomy and Astrophysics, 1900–1975, ed. K. R. Lang & O. Gingerich (Cambridge, MA: Harvard Univ. Press), 716

Ivezić, Ž, Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111

Kartaltepe, J. S., Mozena, M., Kocevski, D., et al. 2015, ApJS, 221, 11

Kim, J., Hur, Y., Park, S., et al. 2020, Advances in Neural Information Processing Systems 33 (NeurIPS 2020) (Online: NIPS), 14567

Laureijs, R., Amiaux, J., Arduini, S., et al. 2011, arXiv:1110.3193

Lee, D., et al. 2013, in Workshop on Challenges in Representation Learning (Atlanta: ICML), 896

Lee, H., Shin, S., & Kim, H. 2021, in Advances in Neural Information Processing Systems 34, ed. M. Ranzato et al., 34 (Online: NIPS), 7082

Li, G., Xu, T., Li, L., et al. 2023, MNRAS, 523, 488

Lotz, J. M., Primack, J., & Madau, P. 2004, AJ, 128, 163

Ma, Z., Zhu, J., Zhu, Y., & Xu, H. 2019, in Int. Conf. Data Mining and Big Data, ed. J. Filipe et al. (Berlin: Springer), 191

Miyazaki, S., Komiyama, Y., Nakaya, H., et al. 2012, Proc. SPIE, 8446

Parry, O., Eke, V., & Frenk, C. 2009, MNRAS, 396, 1972

Peng, C. Y., Ho, L. C., Impey, C. D., & Rix, H.-W. 2002, AJ, 124, 266

Reza, M. 2021, A&C, 37, 100492

Salucci, P. 2019, A&ARv, 27, 1

Slijepcevic, I. V., Scaife, A. M., Walmsley, M., et al. 2022, MNRAS, 514, 2599

Sohn, K., Berthelot, D., Carlini, N., et al. 2020, Advances in Neural Information Processing Systems 33 (Vancouver: NIPS), 596

Soroka, A., Meshcheryakov, A., & Gerasimov, S. 2021, arXiv:2105.02958

Tang, H., Scaife, A. M., & Leahy, J. 2019, MNRAS, 488, 3358

Wei, C., Sohn, K., Mellina, C., Yuille, A., & Yang, F. 2021, in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (Nashville, TN: IEEE), 10857

Wei, S., Li, Y., Lu, W., et al. 2022, PASP, 134, 114508

Wijesinghe, D., Hopkins, A., Kelly, B., Welikala, N., & Connolly, A. 2010, MNRAS, 404, 2077

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, MNRAS, 435, 2835

Wu, D., Zhang, J., Li, X., & Li, H. 2022, RAA, 22, 115011

Xu, Y., Shang, L., Ye, J., et al. 2021, PMLR, 139, 11525

Yang, X., Song, Z., King, I., & Xu, Z. 2022, IEEE Trans. Knowl. Data Eng., 35, 8934

York, D. G., Adelman, J., Anderson, J. E., Jr., et al. 2000, AJ, 120, 1579

Zhang, Z., Zou, Z., Li, N., & Chen, Y. 2022, RAA, 22, 055002

Zhu, X.-P., Dai, J.-M., Bian, C.-J., et al. 2019, Ap&SS, 364, 55

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

Source: ChinaXiv — Machine translation. Verify with original.