IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 February 2024, accepted 22 February 2024, date of publication 26 February 2024, date of current version 4 March 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3369919*

RESEARCH ARTICLE

# Research on Multi-Label Semi-Supervised Learning Algorithm Based on Dual Selection Criteria

**RONGXIN LIU[1], YUFANG LU[1], (Member, IEEE), LEI SHI[2], AND SHURU TAN[1]**
[1]School of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China
[2]School of Science, Guilin University of Technology, Guilin 541004, China

Corresponding author: Yufang Lu (luyufang165@163.com)

This work was supported by the 2022 Guangxi Natural Science Foundation of China under Grant 2022GXNSFAA035584.

**ABSTRACT** With the rapid development of information technology, efficient multi label classification of massive data is one of the important tasks of big data systems. Semi supervised learning algorithm is an effective data classification method, currently mainly applied to the classification of single label data. This article proposes a multi label dynamic semi supervised learning algorithm based on dual selection criteria. The algorithm mainly establishes dual selection criteria for multi label pseudo labeled samples based on the COIN structure and K-nearest neighbor algorithm. A novel pseudo labeled sample selection method is designed, which improves the robustness and accuracy of the algorithm and effectively solves the problem of not considering sample correlation when selecting pseudo labeled samples. On this basis, by adding a performance evaluation mechanism to the model, the model can dynamically and adaptively extract pseudo labeled samples, improving the training speed and accuracy of the model. This article selected four convincing public test datasets for experiments, and the experimental results showed that the proposed semi supervised learning method has improved in multiple indicators such as robustness, accuracy, and training efficiency compared to current mainstream methods.

**INDEX TERMS** Multi-label, semi-supervised learning, COIN structure, K-nearest neighbor algorithm.

## I. INTRODUCTION

In recent years, with the rapid development of information technology and the improvement of information technology [1], the data generated by various fields have been growing geometric progression. Huge amounts of high-dimensional data appear in all aspects of people's lives, such as: Medical Diagnosis, health care, drug development, social media, e-commerce, transportation information, economic and financial services, and online education.. etc [2]. It is quite easy to collect a large number of unlabeled data, but relatively difficult to obtain a large number of labeled data. Therefore, how to use a large number of unlabeled data to improve learning performance has become one of the most concerned problems in machine learning research [3], [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

Semi-supervised learning and self-supervised learning can work with large amounts of unlabeled and labeled data, but in the multi-label classification problem, because of the the large number of labels it is difficult to label the data [5]. In order to overcome this problem, many scholars have studied the semi-supervised multi-label classification problem. How to train a reliable model by partial labeled data and a large number of unlabeled data has become a major problem to be solved. At present, semi-supervised learning mainly focuses on the single label classification field, and there are few studies on multi-label semi-supervised learning. The existing semi-supervised learning has the problems of slow learning speed and single criteria for selecting pseudo-labeled data [6], [7], [8].

In semi-supervised label classification [6], data sets are mainly composed of fully labeled data and unlabeled data. At present, semi-supervised learning mostly serves for single-label problem, because of the particularity of multi-label,

it can not use the prediction probability of single label as the method of selecting pseudo-label data, therefore, semi-supervised learning for multi-label problems is more complex than single-label problems. The current multi-label semi-supervised learning mainly includes the methods based on Graph Model [7], [8] and Double classifier structure [9]. In the study of graph-based methods, Sun et al. [10] proposed semi-supervised low-rank Mapping Learning, forcing classifiers to be low-rank while adding a manifold regularization term to ensure data smoothness [9]. Then Jing et al [11] put forward the multi-label course learning of graph data based on the course learning strategy, and applied the course learning to graph data. After that, in order to solve the problem of the credibility of the icon labels, Wang et al. [8] proposed a writing-based multi-label communication method, which further improved the credibility of labels. Liu proposed an optimal sample selection strategy that can guide the improvement of model performance and is of great help to weakly supervised learning systems based on data [12]. Yang proposed a new framework based on semi supervised multi label deep learning to solve the multi label classification problem of non-invasive load monitoring (NILM), reducing dependence on large label datasets [13]. Wei proposed a Reliable Label Selection and Learning (ReLSL) algorithm to solve the problem of semi supervised deep learning when there is only a very small amount of labeled image data [14]. Tang proposed a u-wordMixup method for data augmentation of unlabeled samples, which solves the quality problems that may arise from unlabeled and annotated samples coming from different fields, and improves generalization ability and accuracy [15]. Wu proposed a Conditional Consistency Regularization (CCR) tailored for Semi Supervised Single Label Image Classification (SS-SLC), which encourages two predictions to remain consistent and establishes a relationship between given two different label states, which helps to utilize label relationships to promote image classification [16]. For different application backgrounds, other similar graph-based methods can be found in literatures [17], [18], [19], [20], and [21]. In the study of methods for dual classifier structure (COIN) [9], Zhan [9] first proposed a COIN-based approach to apply the joint training strategy to multi-label semi-supervised learning.In each joint training, label information on the feature space is learned by maximizing the diversity between the two classifiers on the feature subset. Then, the unlabeled data were predicted by pairwise sorting, and the pseudo-labeled data were selected based on the difference value of dichotomy for iterative training to optimize the model Studying. Based on Zhan, Chu [22] further proposes an integrated approach to accommodate stream-style multi-label data. Then Wang [23] proposed dual-relational semi-supervised multi-label learning (DRML), and designed a bi-classifier domain adaptive network to align features in potential spaces. Li proposed a two-stage training strategy for robust domain adaptation, which effectively utilizes unlabeled target data to generate pseudo labels and pseudo boundaries, thereby achieving model adaptation without the need for source data [24]. Liu introduced fuzzy reasoning into the tracking process to analyze the reliability of the detection graph and improve the robustness of the algorithm [25]. Huang proposed a percentage based threshold adjustment scheme to dynamically change the score thresholds of positive and negative pseudo labels for each category during the training process, as well as the dynamic unlabeled loss weights, thereby further reducing the noise of early unlabeled predictions [26]. Rahman first proposed a new graph convolution based decoder for general semantic and image segmentation tasks [27]. Adiga proposed estimating segmentation uncertainty by utilizing global information from segmentation masks and using a single inference to estimate uncertainty, thereby reducing the total computational cost [28]. Han proposed research directions in areas such as complex concept drift, complex label association, feature selection, and class imbalance [29]. Other similar approaches are described in literatures [25], [30], [31], [32], and [33]. The existing graph-based model reflects the similarity between nodes, while dual classifier structure model solves the problems of insufficient data generalization and complex structure, however, little consideration has been given to the correlation between the data and the long training time. Therefore, it is necessary to consider the relevance of pseudo-marker data selection and the improvement of model training efficiency.

Based on the above analysis, this paper proposes a multi-label dynamic semi-supervised learning method based on double selection criteria (DMSD) It can not only learn label relationships from labeled data, but also extend to unlabeled data. At the same time, a multi-label pseudo-label data selection method based on COIN structure and K-nearest neighbor algorithm is proposed, the problem of not considering the correlation between data is solved. The main contributions of this paper are as follows:

This paper proposes a new method to select pseudo-label data of multi-label based on double classification structure difference value and neighbor difference, which solved problem that the present double classification structure method is not taken the connection between data into account when select pseudo data.

This paper proposes a method to select pseudo-label data based on adaptive difference value. During the training of the model, the number of pseudo-label data can be adjusted adaptively according to the current performance index of the model, the model training efficiency is improved.

## II. MODELING

### A. MODEL STRUCTURE

The basic COIN structure of the model established in this article is based on the COIN method, which applies joint training strategies to multi label semi supervised learning. In each round of joint training, the label information in the feature space is learned by maximizing the diversity between two

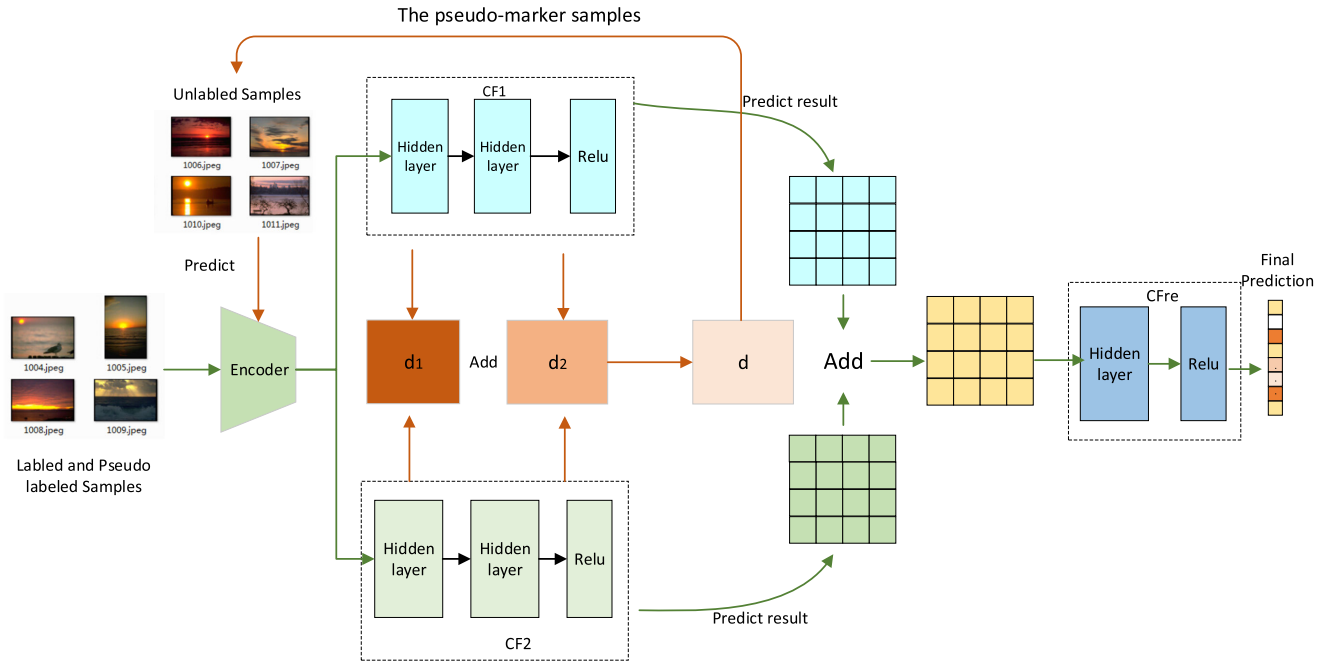


**FIGURE 1.** The DMSD model framework.

classifiers on the binary feature subset. Then, paired sorting prediction is performed on the unlabeled data, and pseudo labeled samples are selected for iterative training based on the difference values of binary classification to optimize the model.

A multi-label dynamic semi-supervised learning method (DMSD) based on double selection criterion is proposed in this paper. The concrete model structure is shown in Figure 1. $X_l$, $Y_l$ represents the given label data. $X_l \in R^{n_l * d}$ represents the eigenvector matrix of the labeled data. $Y_l \in R^{n_l * d_l}$ represents the corresponding label matrix. $n_l$ represents the number of data for the label data. d and $d_l$ represent the feature dimensions of the label data and the number of label categories. The line $x_i \in R^d$ of $X_l$ represents the ith data data. $y_i$ is the label for $x_i$. $X_u \in R^{n_l * d}$ and $Y_u \in R^{n_l * d_l}$ represents an unlabeled feature matrix and a labeled matrix.

d1:Calculating the difference value.

d2:Calculate the difference between the prediction label and its n nearest neighbors.

d:Combined difference value.

The DMSD model structure consists of a feature encoder $E$ , a final prediction network $CF$ , two multi-label classifiers $CF$ , $CF$ . At the beginning of the algorithm, $E$ is used to encode all data into the same feature space.

$$M_l = E(X_l), \tag{1}$$

$$M_u = E(X_u), \tag{2}$$

$M_l \in R^{n_l * d_s}$ and $M_u \in R^{n_u * d_s}$ Is the encoded feature matrix. $d_s$ represents the encoded feature dimension. In the framework of this paper, $CF_1$, $CF_2$ is used to obtain the initial prediction results.so we can get the loss LS1:

$$LS_1(X_l, Y_l) = \frac{1}{2}[\| CF_1(M_l) - Y_l \|_F^2 + \| CF_2(M_l) - Y_l \|_F^2], \tag{3}$$

The optimization objectives is:

$$\min LS_1(X_l, Y_l). \tag{4}$$

Because $X_l$ and $X_u$ can be obtained from a variety of sources, the distribution of the two may be somewhat different, and inspired by the article [34], By predicting the difference between the sample and the original training sample, the reliability of the sample can be further improved. This paper adopts a dual classification structure to achieve consistent distribution of labeled and unlabeled data. In the first part of the training $CF_1$, $CF_2$ and encoder $E(.)$ are trained in an adversarial learning manner to align their distributions (to maximize the variance of the multi-label instance $M_u$), the difference of the two classifiers can be estimated by $l_1$norm:

$$d(p_{1i}, p_{2i}) = \frac{1}{d_l} \| p_{1i} - p_{2i} \|_1, \tag{5}$$

In formula (5) $\| . \|_1$ represents the $L1$ norm calculation. $p_{1i} \in R^{d_l}$ and $p_{2i} \in R^{d_l}$ represents the predicted results of the classifiers $CF_1$, $CF_2$.The $LS_2$ of the second loss function measures the variance of two predictions:

$$LS_2(X_u) = d(CF_1(M_u), CF_2(M_u)), \tag{6}$$

Therefore, the first part of the training objective is to maximize the classification difference, the loss function is as follows:

$$\min -LS_2(X_u) + \lambda LS_1(X_l, Y_l), \tag{7}$$

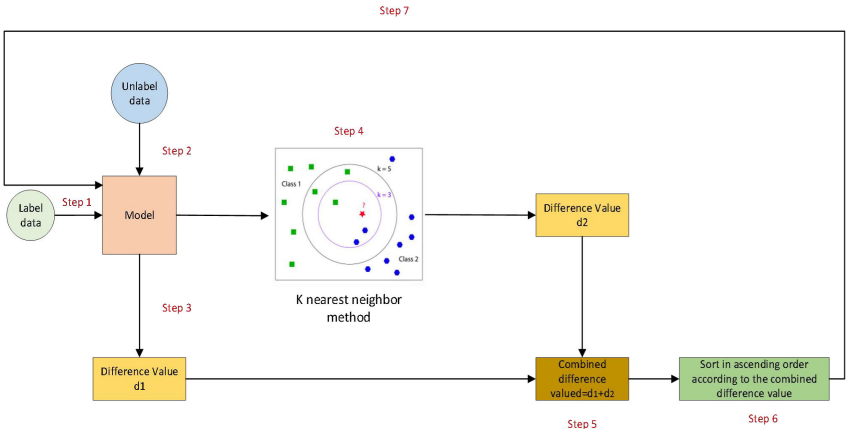


**FIGURE 2.** Pseudo-label data selecting method.

λ is a super-parameter used to control the training weights of the two loss functions.

At the same time, $E$ Needs to learn the coding representation in the subspace to reduce the difference of classification results. Therefore, the objective loss function of updating $E$ is as follows:

$$\min \mathrm{LS}_2, \tag{8}$$

The relationship between labels is an important factor to improve the effectiveness of multi-label recognition. The above-mentioned $CF_1$, $CF_2$ Only made a simple prediction, but did not make use of the relationship between labels. Therefore, this paper will obtain the prediction result matrix, $RE_1 \in R^{n_l * d_l}$ and $\mathrm{RE}_2 \in \mathrm{R}^{\mathrm{n_l * d_l}}$, by arranging the prediction results after $CF_1$, $CF_2$:

$$\mathrm{RE}_1 = [\mathrm{r}_{11}, \mathrm{r}_{12}, \ldots, \mathrm{r}_{1\mathrm{i}}, \ldots \mathrm{r}_{1\mathrm{n_l}}], \tag{9}$$

$$\mathrm{RE}_2 = [\mathrm{r}_{21}, \mathrm{r}_{22}, \ldots, \mathrm{r}_{2\mathrm{i}}, \ldots \mathrm{r}_{2\mathrm{n_l}}], \tag{10}$$

After adding $RE_1$ to $RE_2$, an activation function $\delta$ is used to obtain tensor $F \in R^{n_l * d_l}$ containing label relationship information.

$$\mathrm{F} = \delta(\mathrm{RE}_1 + \mathrm{RE}_2), \tag{11}$$

Finally, the obtained $F$ is handed over to the final neural network $Cre$ For the final prediction of the label, and the predicted loss function is as follows:

$$\mathrm{LS}_{\mathrm{Cre}} = \sum_{\mathrm{i}=1}^{\mathrm{n_l}} \parallel \mathrm{y_i} - \mathrm{CFre(F)} \parallel_2^2, \tag{12}$$

All networks participate in iterative updates during the training, so the final loss function is as follows:

$$\min \frac{\alpha}{2} \mathrm{LS}_1 + (1-\alpha)\mathrm{LS}_{\mathrm{CFre}}. \tag{13}$$

In this paper, the structure of $CF_1$, $CF_2$ is the same, they are small-scale neural networks with 4 layers, including one input layer, two hidden layers and one output layer, the number of neurological source is the number of labels in the data, and the model only contains a Relu activation function after the last hidden layer. $CFre$ is a three-layer linear neural network with only one hidden layer, and the number of neurons is equal to the number of labels in the training data. And the implied layer contains a *Relu* activation function.

### B. SEMI-SUPERVISED LEARNING METHOD

#### 1) DOUBLE PSEUDO-LABEL SELECTION CRITERIA

This article proposes a multi label pseudo labeled sample selection method based on the COIN structure, combined with the K-nearest neighbor algorithm strategy. By combining the difference values of the dual classifiers and the K-nearest neighbor algorithm to select pseudo labeled samples, the problem of not considering the correlation between samples is solved. By predicting the difference between the sample and the original training sample, the reliability of the sample can be further improved.

In single-label semi-supervised training, the prediction results of the model can be directly used to select pseudo-label data, thus evolving a variety of pseudo-label assignment strategies. However, in multi-label learning, because of the label numbers increased label prediction can not be used directly as the basis for the selection of pseudo-label data. Thus, drawing on the idea of a dual classifier architecture [23], this paper proposes a semi-supervised learning approach based on the COIN structure and K-nearest neighbor algorithm to solve the above challenges, with a detailed process as shown in Figure 2. Firstly, the predicted results of $C$ and $C$ are reused, and the difference between them is calculated as the first evaluation index of label selection. The formula for calculating the variance is as follows:

$$\mathrm{dm}_1(\mathrm{x_i}) = \parallel \mathrm{CF}_1(\mathrm{E(x_i)}) - \mathrm{CF}_2(\mathrm{E(x_i)}) \parallel_2^2. \tag{14}$$

Secondly, the K-nearest neighbor algorithm was used to screen the pseudo-marker data, and then the difference value was calculated as the second evaluation index:

$$\mathrm{dm}_2(\mathrm{x_i}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{j}=1}^{\mathrm{N}} \parallel CFre(CF(\mathrm{x_i})) - \mathrm{Y_{ij}} \parallel_2^2. \tag{15}$$

where $N$ is the number of nearest neighbor selection of K nearest neighbor algorithm. $Y_{lj}$ is the actual label value of the $j$ nearest neighbor of data $x_i$ predicted by the model. After obtaining two evaluation indicators, the final pseudo-marker data evaluation indicators formula is as follows

$$\mathrm{dm} = \mathrm{adf}_1\ (\mathrm{x_i}) + (1 - \mathrm{a})\mathrm{df}_2(\mathrm{x_i}). \tag{16}$$

**The meaning represented by the numbers in Figure 2**

1.Train.

2.predict.

3. The difference value d 1 is calculated based on dual structure.

4.Based on K-nearest neighbor algorithm, the difference between predicted label and its N-nearest neighbors is calculated.

5.The final difference value d is calculated by adding the difference values.

6. The pseudo-marker samples added to the training set were selected according to the Order of the final difference values.

7.The model was retrained with a new training set.

Among them, is the coefficient used to control the proportion of two indicators in the difference value of pseudo labeled sample selection. Based on multiple experiments and fitting, the optimal coefficient was selected, and the final value in this article was determined to be 0.8.

### 2) AN ADAPTIVE PSEUDO-MARKER DATA SELECTION METHOD

On the basis of establishing a dual pseudo label selection method, this paper further proposes an adaptive extraction method. The extraction methods proposed by similar models are fixed at 1% -3% each time, while the method proposed in this paper adaptively adjusts the number of unlabeled samples extracted based on the current training accuracy of the model, resulting in a significant improvement in overall efficiency.

The specific selection steps are as follows:

**Step 1**: select the data $m(0 < m < 1)$ from the pseudo-marker data as the training data, and calculate the current pseudo-marker data size as $n = n*(1-m)$ .The value of $m$ is 0.01, which is selected based on the lowest value of samples extracted from other models of the same type, which can effectively verify the superiority of the method proposed in this paper.

**Step 2**: find the pseudo-label data with the current total sort position at m to get the difference value dm, count the number of pseudo-label data less than the difference value $d_m$ and mark it as $x$.

**Step 3**: calculate the ratio of the current pseudo-label data less than the difference value d to the total current pseudo-label data $k = x/n$ If $k < m$, it means that the precision of the current model decreases and the number of pseudo-label data allocated needs to be reduced, The next addition of data was $\frac{3}{4}m$; If $m < k < 2m$, the next allocation ratio is $\frac{1}{2}(m+k)$; If $k > 2m$, then the next distribution ratio is adjusted to $2m$; Otherwise, the next allocation ratio remains the same, continue to allocate according tom, each allocation of pseudo-mark data at least 1.

Among them, when k < m is added, the sample ratio is adjusted to $\frac{3}{4}$m, indicating a decrease in the current data processing accuracy. Reducing the ratio by a quarter can effectively ensure the accuracy of the model; When m < k < 2m is used, it indicates that the current accuracy is high and the proportion can be increased. Taking the middle value not only ensures accuracy but also improves the efficiency of the model in processing data; When k > 2m is reached, it indicates that the pseudo labeled samples with small differences currently have a larger extraction amount than twice the current amount. To ensure accuracy, only twice the amount is extracted.

**Step 4**: repeat steps 1-4 until all unlabeled data have been allocated.

## III. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental environment for this article: CPU: 16 core/GPU, Xeon (R) Platinum 8350C; Memory: 42 GB/GPU; Video memory: 24 GB; Floating point computing power: single precision 27.77 TFLOPS/semi precision 117 Tensor TFLOPS.

In the first part, the number of learning iterations for Corel5K data set, CUB data set, YEST data set, and COCO data set are all 100; in the second part, the number of learning iterations is 5000 for one round of Corel5K data set and CUB data set, 25000 for one round of COCO data set, and 2500 for one round of YEST data set. The learning rate is uniformly 0.000001.

During the experiment, three public test multi-label data sets are used, which are Corel5K data set, CUB data set and Yeast data set.

**The Corel5K data set** [35] is an image data set containing photographs from the Corel CD database. There were 4,500 and 499 data for training and testing, respectively. The total number of label candidates was 260, with an average of 3.40 labels per data.

**The CUB data set** is a data set of bird images involving 200 species of birds. This article used 10,000 labels for training and testing. There were a total of 312 labels, with an average of 31.4 labels per data.

**Yeast data set** [36] is an image data set containing Yeast. 2,417 data were used for training and testing. The average total number of candidate tags per data was 14 and 4.23.

The COCO data set is a large and rich data set for object detection, segmentation, and captioning. This data set is aimed at scene understanding, mainly extracting from complex daily scenes, and calibrating the position of targets in images through precise segmentation. The image includes 9 types of targets, 328000 images, and 250000 labels. So far, there is the largest data set with semantic segmentation, providing 80 categories and over 330000 images, of which 200000 are annotated. The total number of individuals in the entire data set exceeds 1.5 million.

Image features are obtained by using VGG19 [37]. The extracted image features 25088 features per data. VGG was

pretrained on ImageNet and fixed throughout. The baseline models for comparison are as follows:

**Benchmark Model 1**: FastTag [38] is designed to handle noise and incomplete training samples. The design consists of two parts, one part uses missing tag training to learn the tag information, and the other part uses to learn the feature information of the picture, the missing label and the linear projection of prediction are completed respectively in the two parts.

**Benchmark Model 2**: Semantic AutoEncoder(SAE) [39] Linear automatic encoder strategy is used to solve the label prediction problem. The encoder and decoder share the same weight to project the feature space into the label space and then return to the feature space.

**Benchmark Model 3**: Dual Relation Multi-label learning (DRML) [23] a two-classifier structure is proposed to align the distribution offset between labeled and unlabeled samples, and a relational learning network is designed to explore the labeled relationship.

**Benchmark Model 4**: Semi-Supervised Dual Relation Learning(SDRL) [35] the label assignment problem of multi-label pseudo-label samples is solved by using double-classifier structure, and the label tensor relation is used to learn the potential relationship between labels.

FastTag and SAE are supervised learning models, only labeled data are used for training, DRML and SDRL are semi-supervised methods, and labeled and unlabeled samples are combined for training.

In order to evaluate the differences between the models, six evaluation indexes, including accuracy rate, accuracy rate, recall rate, *F1* score, absolute matching rate and Hamming loss, were selected according to the relevant evaluation indexes in literature [40], the higher the accuracy rate, the accuracy rate, the recall rate, the absolute matching rate and the F1 score, the better the Hamming loss.

**TABLE 1. The performance of each model on different data sets.**

| Data set | Method | Absolute matchin rate | Hamming loss | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|---|---|---|
| Core l5K | FastTag | 0.00801 | 0.01253 | 0.98543 | 0.57039 | 0.06701 | 0.09705 |
| | SAE | 0.18837 | 0.01689 | 0.98021 | 0.21847 | 0.13201 | 0.11806 |
| | DRML | 0.00801 | 0.01283 | 0.98588 | 0.18941 | 0.07751 | 0.07725 |
| | SDRL | 0.01603 | 0.01182 | 0.97542 | 0.60141 | 0.26497 | 0.30182 |
| | Ours | **0.04208** | **0.01025** | **0.98643** | **0.6166** | **0.33224** | **0.35659** |
| CUB | FastTag | 0.00251 | 0.08245 | 0.88561 | 0.17451 | 0.03646 | 0.02778 |
| | SAE | 0.03111 | 0.05159 | 0.87763 | 0.19368 | 0.11911 | 0.10121 |
| | DRML | 0.00511 | 0.09833 | 0.89871 | 0.37834 | 0.02238 | 0.03262 |
| | SDRL | 0.00951 | 0.05226 | 0.89085 | 0.33204 | 0.17206 | 0.17692 |
| | Ours | **0.24489** | **0.16578** | **0.89937** | **0.72863** | **0.34015** | **0.3908** |
| Yeast | FastTag | 0.01517 | 0.22807 | 0.73389 | 0.73369 | 0.99981 | 0.80016 |
| | SAE | 0.08275 | 0.11931 | 0.67379 | 0.32815 | 0.32528 | 0.28482 |
| | DRML | 0.01517 | 0.22807 | 0.73527 | 0.74334 | 0.99981 | 0.80644 |
| | SDRL | 0.13793 | 0.13981 | 0.75665 | 0.70404 | 0.55084 | 0.50984 |
| | Ours | **0.17793** | **0.15871** | **0.76857** | **0.74851** | **0.99981** | **0.80644** |
| COCO | FastTag | 0.02953 | 0.02953 | 0.47734 | 0.03146 | 0.28186 | 0.05485 |
| | SAE | 0.1976 | 0.02136 | 0.46894 | 0.03473 | 0.04437 | 0.02741 |
| | DRML | 0.072 | 0.0258 | 0.47735 | 0.07811 | 0.17928 | 0.09241 |
| | SDRL | 0.2966 | 0.1485 | 0.47495 | 0.1388 | 0.1339 | 0.10575 |
| | our | 0.2046 | 0.02134 | 0.47756 | 0.15295 | 0.17295 | 0.17295 |

## IV. LABLE CLASSIFICATION ABILITY ASSESSMENT

**A detailed explanation of the relevant indicators in the above table:**

1. Absolute matching rate refers to the fact that for each sample, the prediction is only considered correct if the predicted value is exactly the same as the true value, which means that as long as there is a difference in the prediction results of a category, it is considered incorrect prediction. Therefore, the larger the value, the higher the accuracy of classification.

2.Hamming Loss measures the proportion of incorrectly predicted labels to the number of labels in all samples. So, for the Hamming Loss, the smaller its value, the better the performance of the model.

3.Accuracy refers to the proportion of correctly predicted sample size in the total sample size. It considers both positive and negative samples, but the disadvantage is that it is not suitable for imbalanced data. Therefore, the larger the value, the higher the accuracy of classification.

4.Precision actually calculates the average accuracy of all samples. For each sample, accuracy is the proportion of correctly predicted labels to the total number of correctly predicted labels by the classifier. Therefore, the larger the value, the higher the accuracy of classification.

5.The recall rate actually calculates the average accuracy of all samples. For each sample, recall is the proportion of correctly predicted labels to the total number of correct labels. Therefore, the larger the value, the higher the accuracy of classification.

6.The F1 value is also calculated as the average F1 value of all samples, which can measure the overall accuracy of the model. Therefore, the larger the value, the better.

Table 1 shows that the model proposed in this article has shown some improvement in comprehensive ability compared to previous models on all four datasets:

1.On a relatively simple Yeast data set, the model proposed in this article has an absolute matching rate that is about 3% higher than the best performing SDRL, an accuracy rate that is about 1% higher than the best performing SDRL, an accuracy rate that is about 0.5% higher than the best performing model SDRL, and is on par with the best performing model in

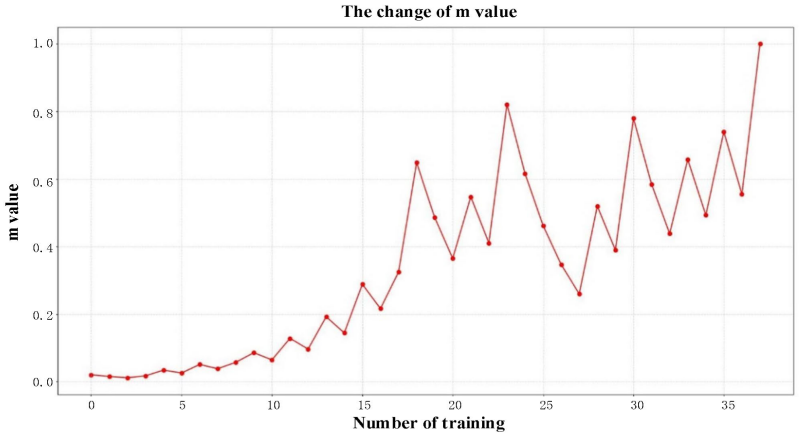


**FIGURE 3.** The change of m value during training.

**TABLE 2.** The training time of each model.

| Method | Training takes time (s) |
|---|---|
| FastTag | 67 |
| SEA | 13147 |
| DRML | 84302 |
| SDRL | 21284 |
| Our | 7251 |

terms of recall and F1 values. In contrast, the model proposed in this article predicts more accurately.

2.On the Corel5k data set, the model proposed in this paper leads the best performing benchmark model by 1% in Hamming loss, 1% higher in accuracy, 1% higher in accuracy, 7% higher in recall, and 5% higher in accuracy. In contrast, the model proposed in this article predicts more accurately.

3.On the most complex CUB data set, the model proposed in this article leads the benchmark model by 20% in absolute matching rate, is basically on par with the best performing model in accuracy, and is 35%, 17%, and 22% ahead of the best performing benchmark model in accuracy, recall, and F1 value. In contrast, the model proposed in this article predicts more accurately.

4.On the COCO data set, the model proposed in this article leads the best performing benchmark model by 10% in accuracy, 17% in F1 value, and slightly ahead of the best performing benchmark model in Hamming loss and accuracy. In contrast, the model proposed in this article predicts more accurately.

From the experiment, it can be seen that the proposed DMSD has better classification performance than previous benchmark models on both simple and complex datasets.

At the same time, as can be seen from figure 3, the overall trend of M values is increasing, which means that the model is indeed improving, therefore, we can also testify the rationality of the adaptive unlabeled sample selection method.

The training time of the model is also an important factor whether the model can be applied in industry. Taking the CUB data set as an example, this paper compares the training time of the benchmark model with that of DMSD, table 2 shows that DMSD outperforms the benchmark model with better classification performance in both classification efficiency and time cost. Compared with the SDRL with the best classification performance in the three data sets, DMSD achieves better performance on CUB data sets with only one-third of the training time of SDRL.

## V. CONCLUSION

The experimental results show that compared with existing benchmark models, our model not only exhibits strong robustness on multiple datasets, but also has significant advantages in training efficiency, which is particularly important when dealing with large-scale datasets.

The multi label semi supervised learning algorithm based on COIN structure and K-nearest neighbor algorithm proposed in this article has achieved significant results in solving the problem of multi label data classification. By introducing the COIN structure and K-nearest neighbor algorithm to establish a DMSD model, we effectively solved the problem of insufficient consideration of sample correlation in multi label learning, which has been a common challenge in previous research. At the same time, based on the DMSD model structure, we have added a new performance evaluation mechanism to enable the model to dynamically and adaptively adjust the amount of pseudo labeled samples extracted. Compared with current mainstream methods, this has improved the training speed and accuracy of the model.

Our research provides new ideas and methods for the field of multi label semi supervised learning. Despite achieving positive results in the experiment, there is still room for further improvement. Future work can be explored in the following directions:

1. Algorithm optimization: We will continue to explore algorithm optimization to further improve the classification accuracy and robustness of the model, especially in handling more complex and noisy data.

2. Ensemble learning strategy: Study how to combine our model with other advanced machine learning technologies, such as ensemble learning, to further enhance the model's generalization ability and performance.

3. Real time learning and online updates: Considering the dynamic and real-time nature of data, future research can focus on how to enable models to learn and adapt to new data streams in real time.

4. Cross domain application: The model proposed in this article performs well on specific datasets, and future research can apply it to a wider range of fields, such as medical diagnosis, financial risk assessment, etc., to verify its applicability and effectiveness in different fields.

5. Interpretability and Transparency: In order to improve the interpretability and transparency of the model, future research can focus on developing methods to explain the decision-making process of the model, which is crucial for understanding and trusting the predictive results of the model.

In summary, this study not only provides new solutions for the field of multi label semi supervised learning, but also lays a solid foundation for future algorithm improvement, application expansion, and theoretical deepening.

## REFERENCES

[1] X. Zhen, "Research on distributed semi supervised learning algorithms," Ph.D. dissertation, School Inf. Electron. Eng., Zhejiang Univ., Hangzhou, China, 2021.
[2] H. Juncheng, "Research on feature selection methods based on multi label learning theory," Ph.D. dissertation, School Comput. Sci. Technol., Jilin Univ., Changchun, China, 2023.
[3] L. Bin, "Research on large-scale semi supervised learning algorithms based on graphs and their applications," Ph.D. dissertation, Univ. Electron. Sci. Technol. China, 2018.
[4] W. Liu, H. Wang, X. Shen, and I. W. Tsang, "The emerging trends of multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7955–7974, Nov. 2022.
[5] W. Hongxin, H. Meng, C. Zhiqiang, Z. Xilong, and L. Muhang, "Multi-label classification under supervised and semi-supervised learning," *Comput. Sci.*, vol. 49, no. 8, pp. 12–25, 2022.
[6] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proc. AAAI*, vol. 6, 2006, pp. 421–426.
[7] X. Kong, M. K. Ng, and Z.-H. Zhou, "Transductive multilabel learning via label set propagation," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 704–719, Mar. 2013.
[8] H. Wang, Z. Li, J. Huang, P. Hui, W. Liu, T. Hu, and G. Chen, "Collaboration based multi-label propagation for fraud detection," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2477–2483.
[9] W. Zhan and M.-L. Zhang, "Inductive semi-supervised multi-label learning with co-training," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 1305–1314.
[10] L. Sun, S. Feng, and G. Lyu, "Robust semi-supervised multi-label learning by triple low-rank regularization," in *Proc. 23rd Pacific-Asia Conf.*, Macau, China, Cham, Switzerland: Springer, Jul. 2019, pp. 269–280.
[11] L. Jing, L. Yang, J. Yu, and M. K. Ng, "Semi-supervised low-rank mapping learning for multi-label classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1483–1491.
[12] S. Liu, X. Xu, Y. Zhang, K. Muhammad, and W. Fu, "A reliable sample selection strategy for weakly supervised visual tracking," *IEEE Trans. Rel.*, vol. 72, no. 1, pp. 15–26, Mar. 2023.
[13] Y. Yang, J. Zhong, W. Li, T. A. Gulliver, and S. Li, "Semisupervised multilabel deep learning based nonintrusive load monitoring in smart grids," *IEEE Trans. Ind. Informat.*, vol. 16, no. 11, pp. 6892–6902, Nov. 2020.
[14] W. Xiang et al., "ReLSL: Semi supervised learning algorithm based on reliable label selection and learning," *J. Comput. Sci.*, vol. 45, no. 6, pp. 1147–1160, 2022.
[15] T. Huanling et al., "A semi supervised deep learning model based on u-wordMixup," *Control Decis. Making*, vol. 38, no. 6, pp. 1646–1652, 2023.
[16] Z. Wu, T. He, X. Xia, J. Yu, X. Shen, and T. Liu, "Conditional consistency regularization for semi-supervised multi-label image classification," *IEEE Trans. Multimedia*, vol. 14, no. 8, pp. 1520–9210, 2023, doi: 10.1109/TMM.2023.3324132.
[17] C. Gong, D. Tao, and J. Yang, "Teaching-to-learn and learning-to-teach for multi-label propagation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, 2016, p. 1.
[18] L. Feng, B. An, and S. He, "Collaboration based multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 3550–3557.
[19] Q. Tan, Y. Yu, G. Yu, and J. Wang, "Semi-supervised multi-label classification using incomplete label information," *Neurocomputing*, vol. 260, pp. 192–202, Oct. 2017.
[20] J. Ding, Q. Song, and L. Jia, "Multi-label k-nearest neighbor classification method based on semi-supervised," *Recent Developments in Mechatronics and Intelligent Robotics*. Cham, Switzerland: Springer, 2019, pp. 544–551.
[21] Y. Liu, F. Nie, and Q. Gao, "Nuclear-norm based semi-supervised multiple labels learning," *Neurocomputing*, vol. 275, pp. 940–947, Jan. 2018.
[22] Z. Chu, P. Li, and X. Hu, "Co-training based on semi-supervised ensemble classification approach for multi-label data stream," in *Proc. IEEE Int. Conf. Big Knowl. (ICBK)*, Nov. 2019, pp. 58–65.
[23] L. Wang, Y. Liu, and C. Qin, "Dual relation semi-supervised multi-label learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 4, 2020, pp. 6227–6234.
[24] L. Li, Y. Zhou, and G. Yang, "Robust source-free domain adaptation for fundus image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 7840–7849.
[25] S. Liu, S. Huang, X. Xu, J. Lloret, and K. Muhammad, "Efficient visual tracking based on fuzzy inference for intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15795–15806, Dec. 2023.
[26] J. Huang, A. Huang, and B. C. Guerra, "Percentmatch: Percentile-based dynamic thresholding for multi-label semi-supervised classification," 2022, *arXiv:2208.13946*.
[27] M. M. Rahman and R. Marculescu, "G-CASCADE: Efficient cascaded graph convolutional decoding for 2D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 7728–7737.
[28] S. Adiga V., J. Dolz, and H. Lombaert, "Anatomically-aware uncertainty for semi-supervised image segmentation," *Med. Image Anal.*, vol. 91, Jan. 2024, Art. no. 103011.
[29] M. Han, H. Wu, Z. Chen, M. Li, and X. Zhang, "A survey of multi-label classification based on supervised and semi-supervised learning," *Int. J. Mach. Learn. Cybern.*, vol. 14, no. 3, pp. 697–724, Mar. 2023.
[30] S. Li and Y. Fu, "Robust multi-label semi-supervised classification," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 27–36.
[31] A. Mittal, N. Sachdeva, S. Agrawal, S. Agarwal, P. Kar, and M. Varma, "ECLARE: Extreme classification with label graph correlations," in *Proc. Web Conf.*, Apr. 2021, pp. 3721–3732.
[32] H. Ye, Z. Chen, and D. H. Wang, "Pretrained generalized autoregressive model with adaptive probabilistic label clusters for extreme multi-label text classification," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2020, pp. 10809–10819.
[33] A. Mittal, K. Dahiya, S. Agrawal, D. Saini, S. Agarwal, P. Kar, and M. Varma, "DECAF: Deep extreme classification with label features," in *Proc. 14th ACM Int. Conf. Web Search Data Mining*, Mar. 2021, pp. 49–57.
[34] L. Wang, Y. Liu, H. Di, C. Qin, G. Sun, and Y. Fu, "Semi-supervised dual relation learning for multi-label classification," *IEEE Trans. Image Process.*, vol. 30, pp. 9125–9135, 2021.
[35] F. Zhao and Y. Guo, "Semi-supervised multi-label learning with incomplete labels," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 4062–4068.
[36] V. Athitsos and S. Sclaroff, "Boosting nearest neighbor classifiers for multiclass recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2005, p. 45.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[38] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1274–1282.
[39] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4447–4456.
[40] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognit.*, vol. 118, Oct. 2021, Art. no. 107965.

**RONGXIN LIU** received the bachelor's degree in automation from Hunan Institute of Technology, in 2017. He is currently pursuing the master's degree in computer science and technology with Guilin University of Technology. His current research interests include artificial intelligence and machine learning.

**YUFANG LU** (Member, IEEE) received the B.S. degree in measurement and control technology and instruments from the University of Electronic Science and Technology of China, Chengdu, China, in 2005, the M.S. degree in electronic and communication engineering from Guilin University of Electronic Technology, Guilin, China, in 2014, and the Ph.D. degree in integrated circuit system design from Xidian University, in 2023. Since 2019, he has been a Professor with Guilin University of Technology, Guilin. His current research interests include smart grid and power system integration, servo control, and motor drive control.

**LEI SHI** received the B.S. and M.S. degrees in mathematics from Yunnan University, Kunming, China, in 2007 and 2010, respectively, and the Ph.D. degree in mathematics from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2020. He is currently a Lecturer with the School of Science, Guilin University of Technology, Guilin, China. His current research interests include chaos synchronization, discontinuous dynamical systems, neural networks, epidemic dynamics, and biomathematics.

**SHURU TAN** is currently pursuing the master's degree with Guilin University of Technology. His main research interests include machine learning, semi-supervised learning, and multi-label learning.