

Robustness of Classification Algorithm in the Face of Label Noise

Jiawei Zhao^{1,*}, Mengyao Kang¹, and Zheng Han¹

¹Faculty of Engineering, The University of Sydney, Sydney, Australia

Abstract

Label noise is an important part in the process of machine learning. Transition matrix provides an effective way to reduce the impact of label noise on classification algorithm. In this experiment, we study logistic regression algorithm and random forest algorithm. We use the known real transition matrix to evaluate the robustness of the algorithm on two datasets. We also design a transition matrix estimator to estimate the transition matrix of three datasets and evaluate the robustness of the two algorithms. We use average error to evaluate the effectiveness of the transition matrix estimator and the top-1 accuracy to evaluate our method.

Keywords: Machine learning, Label noise, Robustness of algorithm, Transition matrix

Received on 21 April 2023, accepted on 24 April 2023, published on 28 June 2023

Copyright © 2023 Jiawei Zhao *et al.*, licensed to EAI. This is an open access article distributed under the terms of the [CC BY-NC-SA 4.0](#), which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetiot.v9i1.3270

1. Introduction

With the continuous development of machine learning technology, more and more fields begin to apply supervised machine learning to help their daily work. Labels are very important for supervised learning. Wrong labels may affect the efficiency of machine learning and lead to serious consequences. However, label noise is usually generated in different ways in real life, such as human subjective bias, insufficient information of label noise, wrong coding, and communication problems (Frenay & Verleysen, 2014). In order to face the label noise that may appear at any time, it is very important to deal with the label noise effectively.

There are usually three methods to deal with label noise. The first method is to use an algorithm that is robust to label noise (Frenay & Verleysen, 2014). This processing method depends on the algorithm that is insensitive to label noise, but it does not fundamentally solve the problem caused by noise. The second method is to improve the quality of training data (Frenay & Verleysen, 2014). This method is to identify the wrong labels in the sample in advance, and then re-label or delete them directly. However, this method may cause greater damage to the original dataset, such as the loss of a large

amount of data. The third method is to model the tag noise directly (Frenay & Verleysen, 2014).

In this study, the process of using the transition matrix lifting algorithm to improve the robustness of label noise belongs to the third method. In this study, we train and test the logistic regression model and random forest model with the given transition matrix and related datasets. Then, two transition matrix estimators are created using the estimation method based on the confusion matrix and the estimation method based on the anchor point and estimated on two label noise datasets with the real transition matrix. We use average error to prove the effectiveness of the estimator. In addition, we also estimate the transition matrix of the CIFAR dataset with unknown flip rates and evaluate our method using the average and standard deviation of the top1 accuracy. The purpose of this study is to evaluate better transition matrix estimation methods and explore the robustness of different classification algorithms to label noise.

2. Previous Work

The earliest discovery of label noise can be traced back to the dichotomy of arbitrary classification noise in the 1980s. In the

*Jiawei ZHAO. Email: zjweiok@gmail.com

process of data set production, label noise will appear in the data set due to various reasons, and label noise will have a certain impact on the performance of the final classification model of the data. Therefore, in the process of machine learning, it is necessary to clean the existing data to avoid the interference of label noise. The advantage of tag noise is that it can be used as data tags in all kinds of data and our real life. If we can label the data with noise, we can quickly find the data we need in a large amount of data. The disadvantage of label noise is also obvious. In machine learning, we need to assume that each labelled data set we get contains noise. Moreover, due to the large sample size, it is impossible to manually check and proofread the labels one by one for each labelled data set. Therefore, we can only predict the data through the machine learning model, and the label noise will have a certain impact on the results of the prediction mode.

The learning methods of label noise are usually divided into two categories: statistical consistent/inconsistent and risk consistent classifiers. The first kind of noise can limit the noise through heuristic methods, so as to reduce the impact of noise on data. For the second kind of noise, we first assume that the label of noise is random noise, and minimize the impact of noise through the loss correction process (1). Where $Q(f_{\theta}(x_i)) = p(y \sim | f_{\theta}(x_i))$. Q can be represented by noise transfer matrix T , so that $Q(f_{\theta}(x_i)) = T f_{\theta}(x_i)$. Where each element of the matrix represents the transition probability from $y \sim$ noise to y is $T_{ij} = p(y \sim = j | y = i)$ (Díaz & Steele, 2021).

$$\hat{R}_{l,D}(f) = \frac{1}{N} \sum_t^N l(Q(f_{\theta}(x_i)), \hat{y}_i). \quad (1)$$

Using anchor points to train the transfer matrix is a good method to reduce the influence of noise, and can make the effect of classifier more stable. For example, given an instance x , if $P(Y=i | X = x) \approx 1$ and otherwise $P(Y = k | X = x) = 1$ where $k \neq i$, we will be able to rely on this information to obtain the transition matrix (Díaz & Steele, 2021). However, in the case of no anchor point, the transfer matrix will be incorrectly learned, which will increase the impact of label noise on the data and greatly reduce the stability of the classifier.

$$P(\bar{Y} = j | X = x) = \sum_1^c T_{kj} P(Y = k | X = x) = T_{ij}. \quad (2)$$

The method of sample importance weighting is also a good method to improve the effectiveness of training. It assigns weights to instances according to the noise level, and different levels of noise will be assigned different weights to reduce the impact of low-level noise on data. However, this method will excessively rely on the actual weight. When weighting the noise, the correct mark will make the instance have a larger value, and the wrong mark will make the instance smaller and more difficult to predict.

3. Label Noise Methods with Known Flip Rates

When facing the noise label task with known flip rate, we can directly use the given transition matrix to create a classification model robust to label noise. In this experiment, we use logistic regression classifier and random forest classifier as the basis of classification model and use training set to train the classifier, so that the classifier matches the training set with label noise. When using clean test set data for prediction, first use the trained classifier to output the probability of noise interference that classifies each sample to different labels. Then, the given transition matrix is multiplied by these probabilities to obtain the modified classification probability, and finally the samples are assigned to the category with the highest probability. The following describes the two basic classifiers used in building the model.

3.1. Logistic Regression

Logistic regression is a common machine learning framework, which is usually used as a classification model to create binary classification. The logistic regression model can usually be represented by (3), where x represents the characteristics of the input model and θ represents the weight assigned by the model to different characteristics (Hou et al., 2014). By adding a dummy variable x_0 with a value of 1 to (3), Formula 1 can be simplified to (4) (Hou et al., 2014). In continuous iterative learning, the logistic regression model will constantly update the weight and output the final weight allocation scheme for prediction.

$$h(x) = h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n \quad (3)$$

$$h_{\theta}(x) = \theta^T X \quad (4)$$

The prediction method of logistic regression model is to calculate the probability that the sample belongs to a label. Its calculation formula is based on sigmoid function and expressed by (5) (Hou et al., 2014).

$$P(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}} \quad (4)$$

In the task of multiple classification, logistic regression model usually transforms multiple classification tasks into multiple binary classification tasks. The model uses each label as a positive class in turn and the other labels as negative classes to calculate the probability of assigning samples to each positive class label, and finally selects the label with the highest probability as the final classification.

3.2. Random Forest

Random forest is an integrated algorithm, which is based on decision tree. By combining the classification results of

multiple individual decision trees, the final classification results are obtained by voting or taking the mean. Each decision tree is completely independent, they are not related to each other, and the prediction of each tree will not affect the prediction of other trees. The classification accuracy of this algorithm is often higher than that of a single decision tree algorithm, and it can show better robustness in the face of noise. Some previous studies have shown that the random forest classifier itself has certain robustness to label noise, but it still cannot deal with too much or too complex noise (Maas & Heipke, 2019).

4. Noise Rate Estimation Method

The noise rate represents the probability that each label is assigned a different label after adding noise to the labels of the data set. In our research, we use the data set with label noise to iteratively train the specific classifier, and use the classification effect of the specific classifier after learning the noise data set to estimate the noise rate of the label. The noise rates of all category labels will be summarized in a matrix to form a transition matrix (T_i). \hat{Y} in the transition matrix represents the noise label and Y represents the real label. Each column of the transition matrix represents the probability that a certain type of label is assigned to different labels after noise pollution (6).

$$T_i = \begin{bmatrix} P(\hat{Y} = 0 | Y = 0) & P(\hat{Y} = 0 | Y = 1) & P(\hat{Y} = 0 | Y = 2) \\ P(\hat{Y} = 1 | Y = 0) & P(\hat{Y} = 1 | Y = 1) & P(\hat{Y} = 1 | Y = 2) \\ P(\hat{Y} = 2 | Y = 0) & P(\hat{Y} = 2 | Y = 1) & P(\hat{Y} = 2 | Y = 2) \end{bmatrix} \quad (6)$$

4.1. Method Based on Confusion Matrix

In the past research on supervised machine learning models, we usually use confusion matrix to evaluate the classification effect of classifier. For the three classification labels similar to those in this experiment, Y_{pre} in the confusion matrix (T_c) represents the predicted label and Y_{true} represents the real label (7). Each row of the confusion matrix represents the result that a certain type of label is assigned to different labels in the allocation process. A new probability matrix (T_{cp}) can be obtained by calculating the proportion of each value in the row in which the value is located (8). Each row of the probability matrix (T_{cp}) represents the result that a certain type of label is assigned to different labels in the allocation process. This is highly similar to the data represented by each column in the transition matrix. Therefore, we think that we can construct the confusion matrix generated by the classifier when training on the noisy data set, and then exchange the rows and columns of the confusion matrix to obtain the final transition matrix.

$$T_c = \begin{bmatrix} \sum(Y_{pre} = 0 | Y_{true} = 0) & \sum(Y_{pre} = 1 | Y_{true} = 0) & \sum(Y_{pre} = 2 | Y_{true} = 0) \\ \sum(Y_{pre} = 0 | Y_{true} = 1) & \sum(Y_{pre} = 1 | Y_{true} = 1) & \sum(Y_{pre} = 2 | Y_{true} = 1) \\ \sum(Y_{pre} = 0 | Y_{true} = 2) & \sum(Y_{pre} = 1 | Y_{true} = 2) & \sum(Y_{pre} = 2 | Y_{true} = 2) \end{bmatrix} \quad (7)$$

$$T_{cp} = \begin{bmatrix} P(Y_{pre} = 0 | Y_{true} = 0) & P(Y_{pre} = 1 | Y_{true} = 0) & P(Y_{pre} = 2 | Y_{true} = 0) \\ P(Y_{pre} = 0 | Y_{true} = 1) & P(Y_{pre} = 1 | Y_{true} = 1) & P(Y_{pre} = 2 | Y_{true} = 1) \\ P(Y_{pre} = 0 | Y_{true} = 2) & P(Y_{pre} = 1 | Y_{true} = 2) & P(Y_{pre} = 2 | Y_{true} = 2) \end{bmatrix} \quad (8)$$

This transition matrix estimation method is based on logistic regression classifier. In order to verify the correlation between training accuracy and label noise, we tested it with clean test set data in FashionMINIST0.3 dataset and data with label noise (the test is a separate experiment, and the test set data is only used to illustrate the availability of this transition matrix estimation method, and the test set data training model is not used in the main experiment of this study). Since the test set data has only 3000 samples, we randomly selected 3060 samples (17%) from 18000 dataset samples with label noise. When using the default maximum number of iterations (100 iterations), we conducted the test 10 times, and took the average value and standard deviation of all results for analysis (Figure 1).

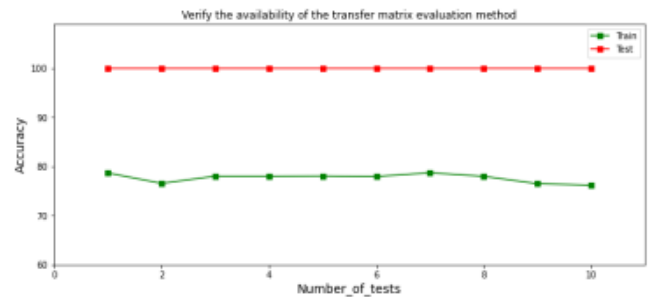


Figure 1. Training accuracy gap of data with label noise

Finally, the classification accuracy of logistic regression classifier on clean datasets is stable at 100%, while for datasets with label noise, the classification accuracy is always less than 80%, and the average value is 77.6503%. This means that the label noise will cause the logistic regression classifier to make wrong judgment on the label of the training set in the process of iterative training. The specific interference caused by label noise to the classification results can be estimated by using the probability matrix (T_{cp}) generated at this stage (8). That is, this method of evaluating the transition matrix is available.

4.2. Method Based on Anchor Point

In our research on machine learning model, we use the anchor transition matrix to evaluate the method. We first use the training model to infer the probability of the model on the training set, and then use the obtained probability to predict the training model. From the predicted probability, we find the anchor data of each class.

Anchor point means that, for example, x_1 belongs to Class1 data, then the probability of $P(Y = 0 | X = x_1)$ is equal to 1, so we can estimate the transfer matrix by this method (9). After finding these data, we find the classification probability in the noise data to find the second anchor point.

From the definition, the second anchor must also belong to Class1. So, we get the probability of the Class1 anchor. Because the anchor point cannot be obtained directly, we need to know it's a priori probability. However, this is a dataset, and we cannot know the a priori probability of the dataset, so we assume that the data most likely to be divided into Class0, and we use it as an anchor. That is, the classification probability of the training set is inversely calculated to obtain the transition matrix of the anchor point.

$$P(Y = i | X = x) = 1 \tag{9}$$

After obtaining the transition matrix of the anchor point, because we are looking for data by taking the column as the object, the data stored in the data is stored in the form of column. After that, we need to transpose the matrix, change the data column into row, and then obtain the transition matrix.

5. Label Noise Methods with Unknown Flip Rates

When facing the noise label task with unknown flip rate, we need to use a specific noise rate evaluation method to estimate the transition matrix, and then use the obtained transition matrix to create a classification model robust to label noise. In this experiment, we use the logistic regression classifier as the basis for estimating the transition matrix and use the training set to train the classifier to match the classifier with the training set with label noise. The flip rate is estimated by using the training results, and the transition matrix is generated. In the test, the transition matrix is used to correct the prediction results, and finally the samples are assigned to the category with the highest probability.

6. Noise Rate Estimation Method

6.1. Datasets

The three datasets used in this experiment are image datasets, which are divided into two basic parts: training and verification data, and test data. The training and validation data correspond to noisy labels with class label noise, and the test data correspond to clean labels. The class set of all labels is {0,1,2}.

FashionMINIST0.3

The number of training and verification data samples in FashionMINIST0.3 dataset is 18000, and the number of test data samples is 3000, in which the shape of each sample image is (28*28). The data set has provided a transition matrix, which can be directly used to design classifiers.

FashionMINIST0.6

The number of training and verification data samples in FashionMINIST0.6 dataset is 18000, and the number of test

data samples is 3000, in which the shape of each sample image is (28*28). The data set has provided a transition matrix, which can be directly used to design classifiers.

CIFAR

The number of training and verification data samples in CIFAR data set is 15000 and the number of test data samples is 3000, in which the shape of each sample image is (32*32*3). The data set does not provide a transition matrix, so it is necessary to use the transition matrix evaluator to estimate the transition matrix, and then use it to design a classifier.

6.2. Performance Evaluation

The Top-1 Accuracy

In order to evaluate the classification performance of different classifiers on three data sets, we use the top1 accuracy metric to evaluate the classification results (10). This index shows the proportion of the number of correctly classified samples in all samples. The larger its value, the better the classification effect of the classifier.

$$\text{The Top-1 Accuracy} = \frac{\text{Number of correctly classified examples}}{\text{Total number of test examples}} * 100 \tag{10}$$

Average Error

In order to explore the effectiveness of the transition matrix estimator created in this experiment, we will use the transition matrix estimator on FashionMINIST0.3 and FashionMINIST0.6 datasets and judge the estimation effect of the transition matrix estimator by determining the gap between the estimated matrix and the given real matrix. The method used to evaluate the gap is to establish a gap matrix (T_{diff}), and each point in the gap matrix represents the absolute value of the value difference at the same position in the estimated transition matrix (T_{es}) and the real transition matrix (T_{true}) (11). The sum of all values in the gap matrix (SUM_{diff}) is then calculated and divided by the number of elements in the gap matrix (N_{diff}). The final value is the average error of all elements in the two matrices (12). The smaller the value, the closer the estimation matrix is to the real matrix.

$$T_{diff} = |T_{es} - T_{true}| \tag{11}$$

$$\text{Average error} = \frac{SUM_{diff}}{N_{diff}} \tag{12}$$

6.3. Experimental Reliability

In order to evaluate the stability of the classifier, this experiment does not use all samples when training the classifier, but randomly selects 20% of the data from the data set for verification and uses the remaining data for training. in addition, in order to ensure that the experimental results are not affected by accidental factors, this experiment repeats ten tests on each data set and classifier, records the top1 accuracy

of ten tests, and calculates the average and standard deviation of these values to evaluate the experimental results.

6.4. Noise Rate Estimation

In order to ensure the reliability of the experiment, each group of experiments will be repeated ten times. Because the sampling process of training data is random, the logistic regression classifier will get different classification results in ten experiments. The method to evaluate the noise rate is based on the classification effect of logistic regression classifier training on noise data, so we will get ten different transition matrices in each group of experiments. We recorded the change of classifier accuracy after using the transition matrix in each experiment and selected the transition matrix with the best change effect on the classifier as the final transition matrix.

7. Results and Analysis

The results and analysis stage are divided into two basic parts, including the experiment of known flip rate and the experiment of unknown flip rate. The experiment of unknown flip rate is divided into two subparts according to different transition matrix evaluation methods, including transition matrix estimation experiment based on confusion matrix and transition matrix estimation experiment based on anchor.

7.1. Known Flip Rates

In the experiment with known flip rate, we directly use the given transition matrix to correct the prediction results of the classifier and show the average top1 accuracy of the experiment in a table. Where ‘LF’ represents logistic regression algorithm and ‘RF’ represents random forest algorithm (Table 1).

Table 1. Classification accuracy using real matrix

| Mean Accuracy (%) | FashionMINIST0.3 | | FashionMINIST0.6 | |
|-----------------------|------------------|---------|------------------|---------|
| | LF | RF | LF | RF |
| Validation | 64.4444 | 68.3806 | 35.2278 | 35.8472 |
| Test without T | 87.3900 | 96.4800 | 54.2700 | 63.1133 |
| Test with T | 93.4567 | 98.1733 | 54.2700 | 62.9900 |

By comparing the results of ‘Validation’ and ‘Test without T’ in Table 1, it can be found that better classification results can be obtained by using the trained model on a clean test set without using the transition matrix. The change degree of accuracy of different models is related to the robustness of the basic classification algorithm used in the model to label noise.

By comparing the prediction results of the two classification algorithms on the test set, it can be seen that the random forest algorithm is more robust than the logistic regression algorithm in the face of label noise. In addition, by comparing the results of ‘Test without T’ and ‘Test with T’ in Table 1, it can be found that in FashionMINIST0.3 data set, the prediction accuracy of the two classification algorithms can be improved by using the transition matrix, and the robustness of the two models has been improved. However, in FashionMINIST0.6 data set, the classification accuracy of the two classification algorithms has not changed significantly, and the accuracy of the random forest algorithm still shows signs of decline. This shows that the transition matrix of FashionMINIST0.6 does not improve the robustness of the classification algorithm. The reason for this may be that FashionMINIST0.6 dataset is more complex, and each category is similar, which will make it difficult to correct the deviation caused by label noise. The average standard deviation of the test results was also recorded (Table 2).

Table 2. Classification standard deviation using real matrix

| Standard Deviation | FashionMINIST0.3 | | FashionMINIST0.6 | |
|-----------------------|------------------|--------|------------------|--------|
| | LF | RF | LF | RF |
| Validation | 0.7150 | 0.5611 | 0.5772 | 0.3573 |
| Test without T | 0.7785 | 0.1284 | 0.7899 | 0.7590 |
| Test with T | 0.5313 | 0.1604 | 0.7899 | 0.7646 |

By comparing the standard deviation of the classification effect of the two classification algorithms on the two datasets in Table 2, it can be seen that the random forest algorithm shows better stability than the logistic regression algorithm in the process of verification and testing. Based on the above results, it can be concluded that the improvement effect of transition matrix on the robustness of classification algorithm will be affected by the complexity of label noise. In addition, we can also find that the random forest algorithm is more robust to label noise when the flip rate is known.

7.2 Unknown Flip Rates

When the flip rate is unknown, we need to estimate the transition matrix first and use the transition matrix to correct the interference caused by label noise. We will first evaluate the transition matrix obtained by the estimator, and then analyse the results obtained by applying the transition matrix.

Method Based on Confusion Matrix

We use the confusion matrix obtained by the logistic regression algorithm in the training process to estimate the transition matrix. The availability of this method has been explained in the previous paper. In this experiment, the transition matrix estimator based on confusion matrix is used

in FashionMINIST0.3 dataset, FashionMINIST0.6 dataset, and CIFAR dataset. In addition, because FashionMINIST0.3 and FashionMINIST0.6 datasets have provided the real transition matrix, we compare the obtained transition matrix

with the real transition matrix. The gap between the two matrices is shown with gap matrix and average error to further illustrate the availability of the estimator (Table 3)

Table 3. Estimation results of transition matrix based on confusion matrix

| | FashionMINIST0.3 | | | FashionMINIST0.6 | | | CIFAR | | |
|-------------------------|------------------|--------|--------|------------------|--------|--------|----------|--------|--------|
| Real matrix | [0.7000 | 0.3000 | 0.0000 | [0.4000 | 0.3000 | 0.3000 | [0.3000 | 0.4000 | 0.3000 |
| | 0.3000 | 0.0000 | 0.7000 | 0.3000 | 0.3000 | 0.4000 | 0.3000 | 0.4000 | |
| Estimated matrix | [0.7129 | 0.2595 | 0.0876 | [0.4457 | 0.2795 | 0.2780 | [0.4457 | 0.2795 | 0.2780 |
| | 0.0448 | 0.6839 | 0.2663 | 0.2865 | 0.6839 | 0.2663 | 0.2865 | 0.4602 | 0.2753 |
| | 0.2423 | 0.0566 | 0.6461 | 0.2678 | 0.2602 | 0.4467 | 0.2678 | 0.2602 | 0.4467 |
| Gap matrix | [0.0129 | 0.0405 | 0.0876 | [0.0457 | 0.0205 | 0.0220 | [0.0135 | 0.0602 | 0.0247 |
| | 0.0048 | 0.0160 | 0.0337 | 0.0135 | 0.0602 | 0.0247 | 0.0577 | 0.0566 | 0.0539 |
| | 0.0577 | 0.0566 | 0.0539 | 0.0322 | 0.0398 | 0.0467 | | | |
| Average error | 0.0449 | | | 0.0339 | | | | | |

By looking at the gap matrix related to FashionMINIST0.3 dataset and FashionMINIST0.6 dataset, we can find that all values in the gap matrix are less than 0.1, and the average error of the two datasets is less than 0.05. This shows that there is little difference between the transition matrix obtained by using the transition matrix estimator we designed and the real matrix. This also shows that the estimator can

make the estimated transition matrix close to the real matrix, and the availability of the transition matrix for CIFAR dataset obtained by the estimator can be further confirmed. We use the evaluated transition matrix to test on three datasets and two classification algorithms. The average accuracy results are shown in the table below (Table 4).

Table 4. Classification accuracy using transition matrix based on confusion matrix

| Mean Accuracy (%) | FashionMINIST0.3 | | FashionMINIST0.6 | | CIFAR | |
|-----------------------|------------------|---------|------------------|---------|---------|---------|
| | LF | RF | LF | RF | LF | RF |
| Validation | 64.2111 | 68.5111 | 35.4556 | 35.4139 | 34.6533 | 35.5733 |
| Test without T | 87.0433 | 96.4633 | 55.5933 | 62.5533 | 41.1400 | 48.7700 |
| Test with T | 91.6567 | 97.8700 | 55.8333 | 58.9733 | 41.8500 | 42.4700 |

Through the results of previous experiments, we know that the robustness of the basic classification algorithm makes the classification model obtain higher accuracy when tested with a clean test set. In the results of this experiment, this conclusion is verified again. In addition, the random forest algorithm itself is more powerful. Therefore, on the three datasets, whether the transition matrix is used or not, the accuracy of the random forest algorithm is higher than that of the logistic regression algorithm. In this experiment, I focus on the performance of the new transition matrix in the experiment. The experimental results show that after using the new transition matrix, the accuracy of the logistic regression algorithm on the three datasets is improved compared with that without the transfer matrix. The accuracy of random forest algorithm has only been slightly improved in the first dataset and decreased in the other two datasets. I

think this may be because we use the logistic regression algorithm as the basic algorithm when generating a new transition matrix. When applying the confusion matrix of the logistic regression algorithm in the training stage, in addition to the label noise, there are some interference factors related to the characteristics of the logistic regression algorithm, which makes the transition matrix more consistent with the logistic regression algorithm. The additional interference factors in the transformation matrix leads to the decline of the accuracy of the random forest algorithm. Nevertheless, the random forest algorithm still gets better classification results by virtue of its good robustness to label noise. In the classification task of CIFAR dataset, the random forest algorithm obtains the highest accuracy of 48.77%, which is still higher than that of logistic regression algorithm even when the interference decreases to 42.47%. The following

figure shows the experiment from the perspective of standard error (Table 5).

Table 5. Classification standard deviation using transition matrix based on confusion matrix

| Standard Deviation | FashionMINIST0.3 | | FashionMINIST0.3 | | CIFAR | |
|--------------------|------------------|--------|------------------|--------|--------|--------|
| | LF | RF | LF | RF | LF | RF |
| Validation | 0.6185 | 0.6663 | 0.8872 | 0.6885 | 0.5534 | 0.5731 |
| Test without T | 0.5721 | 0.1567 | 1.0155 | 1.1483 | 0.7333 | 0.6371 |
| Test with T | 0.4563 | 0.1027 | 0.7729 | 1.2107 | 0.7455 | 0.7130 |

By comparing the standard deviation of accuracy when testing on different datasets in Table 5, we can find that the two classifiers have high stability on FashionMINIST0.3 dataset. This is because the dataset is relatively simple and contains relatively weak noise.

Method Based on Anchor Point

We use the logistic regression algorithm to find the anchor point by assigning the probability of label to each sample in the training process and using the anchor point to estimate the

transition matrix. The details of this method have been described above. This experiment uses anchor-based transition matrix estimator in FashionMINIST0.3 dataset, FashionMINIST0.6 dataset and CIFAR dataset. In addition, because FashionMINIST0.3 dataset and FashionMINIST0.6 dataset have provided the real transition matrix, we compare the obtained transition matrix with the real transition matrix. The gap between the two matrices is shown through the gap matrix and average error to further illustrate the availability of the estimator (Table 6).

Table 6. Estimation results of transition matrix based on anchor point

| | FashionMINIST0.3 | | | FashionMINIST0.6 | | | CIFAR | | | |
|-------------------------|------------------|--------|--------|------------------|--------|--------|--------|--------|--------|--------|
| | | | | | | | | | | |
| Real matrix | [| 0.7000 | 0.3000 | 0.0000 | 0.4000 | 0.3000 | 0.3000 | | | |
| | | 0.0000 | 0.7000 | 0.3000 | 0.3000 | 0.4000 | 0.3000 | | | |
| | | 0.3000 | 0.0000 | 0.7000 | 0.3000 | 0.3000 | 0.4000 | | | |
| Estimated matrix | [| 0.9821 | 0.0309 | 0.0039 | 0.8639 | 0.0807 | 0.0492 | 0.9464 | 0.0985 | 0.0394 |
| | | 0.0152 | 0.9622 | 0.0000 | 0.0803 | 0.8865 | 0.0448 | 0.0374 | 0.8746 | 0.0681 |
| | | 0.0027 | 0.0068 | 0.9967 | 0.0559 | 0.0328 | 0.9060 | 0.0162 | 0.0269 | 0.8925 |
| Gap matrix | [| 0.2821 | 0.2691 | 0.0033 | 0.4639 | 0.2193 | 0.2508 | | | |
| | | 0.0152 | 0.2622 | 0.0300 | 0.2197 | 0.4865 | 0.2552 | | | |
| | | 0.2973 | 0.0068 | 0.2967 | 0.2441 | 0.2672 | 0.5060 | | | |
| Average error | | 0.1925 | | | 0.3236 | | | | | |

By looking at the gap matrix related to FashionMINIST0.3 dataset and FashionMINIST0.6 dataset, we can find that the values in the gap matrix are unevenly distributed, with the maximum difference of 0.4865 and the minimum difference of 0.0033. The average errors of the two datasets are 0.1925 and 0.3236 respectively. This shows that there is a great difference between the transition matrix obtained by using the transition matrix estimator we designed and the real matrix.

This means that the transition matrix obtained by the estimator may not achieve the same effect as the real transition matrix. Therefore, the transfer matrix for CIFAR dataset obtained by this evaluator may not achieve the best performance. We use the evaluated transfer matrix to test on three data sets and two classification algorithms. The average accuracy results are shown in the table below (Table 7).

Table 7. Classification accuracy using transition matrix based on anchor point

| Mean Accuracy (%) | FashionMINIST0.3 | | FashionMINIST0.3 | | CIFAR | |
|-----------------------|------------------|---------|------------------|---------|---------|---------|
| | LF | RF | LF | RF | LF | RF |
| Validation | 64.4889 | 68.5528 | 35.2361 | 35.5583 | 34.9000 | 35.4967 |
| Test without T | 87.3300 | 96.5067 | 55.4011 | 62.4867 | 40.8267 | 49.1600 |
| Test with T | 87.2967 | 96.4500 | 55.5200 | 62.3633 | 41.1993 | 47.1033 |

The experimental results show that after using the transition matrix, the logistic regression algorithm has slightly improved on FashionMINIST0.6 dataset and CIFAR dataset, but slightly decreased on FashionMINIST0.3 dataset. By looking at the data in Tables 1 and Table 4, it can be found that the transition matrix can significantly improve the robustness of the logistic regression algorithm on FashionMINIST0.3 dataset, but not on other datasets. Therefore, when the gap between the estimated transition matrix and the real transition matrix increases, the decline of robustness is more obvious on FashionMINIST0.3 dataset, which may also be related to the simplicity of FashionMINIST0.3 dataset. In addition, we can also notice from Table 7 that the classification accuracy of the random forest classifier on the three datasets has declined to a certain extent when

using the transition matrix estimated according to the anchor point. I think this is still because the transition matrix estimator is based on the logistic regression algorithm, and the additional interference leads to the insufficient quality of the transition matrix. Comparing the performance of random forest algorithm in fashionminist0.6 data set and CIFAR dataset in Table 4 and Table 7, we can see that when using the evaluation matrix based on confusion matrix, the accuracy of random forest algorithm on the two datasets is 58.9633% and 42.4700%. When using the anchor-based evaluation matrix, the random forest algorithm suffers less interference, and the accuracy is improved to 62.3633% and 47.1033%. In addition, we can see from Table 8 that the two classifiers are still the most stable on FashionMINIST0.3 dataset (Table 8).

Table 8. Classification standard deviation using transition matrix based on anchor point

| Standard Deviation | FashionMINIST0.3 | | FashionMINIST0.6 | | CIFAR | |
|-----------------------|------------------|--------|------------------|--------|--------|--------|
| | LF | RF | LF | RF | LF | RF |
| Validation | 0.6813 | 0.5348 | 0.7776 | 0.7560 | 0.5457 | 0.6158 |
| Test without T | 0.5001 | 0.2299 | 1.2830 | 0.7902 | 0.7542 | 1.0983 |
| Test with T | 0.5926 | 0.2040 | 1.2617 | 0.8309 | 0.7498 | 1.2929 |

8. Conclusions

In this study, we use two different classification algorithms as the basic algorithm to construct two models, which use the transition matrix to improve the robustness to label noise. In addition, two different transition matrix estimators are constructed by using confusion matrix and anchor point. Through testing, we evaluate the performance of the model and verify the effectiveness of the method.

In the experiment with known flip rate, we applied the real transition matrix on logistic regression and random forest classifier and tested it on FashionMINIST0.3 dataset and FashionMINIST0.6 dataset. Through the experimental results, we get that on a relatively simple dataset, the

transition matrix does improve the robustness of the model to label noise. In addition, we also know from experiments that the model based on random forest algorithm is more robust to label noise than logistic regression algorithm.

In the experiment of unknown flip rate, we evaluated two different evaluators. We use the mean error to determine the gap between the transition matrix generated by the estimator and the real matrix. The results show that the estimator based on confusion matrix can generate a transition matrix more similar to the real matrix. However, with the progress of the experiment, we found that the more similar transition matrix generated by the confusion matrix estimator only provides better classification accuracy on FashionMINIST0.3 dataset. On other datasets, they interfere with the experimental results. In addition, the two evaluators used in the study are created based on logistic regression algorithm, so the generated transition matrix

may contain other interference factors other than label noise. These factors are related to the logistic regression algorithm, which leads to the improvement of the robustness of the logistic regression algorithm to label noise after using these transition matrices. However, these matrices lead to the decline of classification accuracy of random forest algorithm in the face of label noise. However, due to its strong robustness to label noise, random forest algorithm still shows stronger robustness than logistic regression algorithm after the accuracy decreases. Finally, the random forest algorithm achieves the highest accuracy of 49.16% on CIFAR datasets without real transition matrix. Even after interference, the random forest algorithm still maintains an accuracy of 47.1033%.

9. Future Work

The two transition matrix evaluators established in this study are based on logistic regression algorithm, which makes the transition matrix estimated in this experiment perform poorly on random forest classifier. In addition, the basic classification algorithms used in this study use the default parameters provided by 'Sklearn' library, which may also be one of the reasons for the poor classification accuracy. Based on the above problems, we can continue to expand the research from the following aspects in the future:

- More methods are used to construct the transition matrix estimator, so that the estimated transition matrix can be more accurate and suitable for most classification algorithms.
- By setting different parameters for the classification algorithm, this paper explores the higher accuracy that the classification algorithm can achieve.
- Use more different classification algorithms to explore the robustness of more classifiers under label noise.

References

- [1] Algan, G., & Ulusoy, I. (2020). Label noise types and their effects on deep learning. *arXiv preprint arXiv:2003.10471*.
- [2] Díaz, A., & Steele, D. (2021). Analysis of classifiers robust to noisy labels. *arXiv preprint arXiv:2106.00274*.
- [3] Frénay, B., & Verleysen, M. (2013). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5), 845-869.
- [4] Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., ... & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.
- [5] Hou, T., Zheng, G., Zhang, P., Jia, J., Li, J., Xie, L., Wei, C., & Li, Y. (2014). LAcE: lysine acetylation site prediction using logistic regression classifiers. *PLoS one*, 9(2), e89575.
- [6] Maas, A. E., Rottensteiner, F., & Heipke, C. (2019). A label noise tolerant random forest for the classification of remote sensing data based on outdated maps for training. *Computer Vision and Image Understanding*, 188, 102782.