Combining Outlier Detection and Reconstruction Error Minimization for Label Noise Reduction

Weining Zhang
Department of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, China
zwn7900@nuaa.edu.cn

Xiaoyang Tan
Department of Computer Science and Technology
Nanjing University of Aeronautics and Astronautics
Nanjing, China
x.tan@nuaa.edu.cn

***Abstract*—Label noise is a common phenomenon when labeling a large-scale dataset for supervised learning. Outlier detection is a recently proposed method to handle this issue by treating the outliers of each class as potential data points with label noise and remove them before training. However, this approach could lead to high false positive rate and hurt the performance. In this paper, we propose a novel and effective method to deal with this issue by combining the strength of outlier detection and reconstruction error minimization (REM). The main idea is add a second verification step (i.e., REM) to the outputs of outlier detection so as to reduce the risk of discarding those points which do not fit the underlying data distribution well but with correct label. Particularly, we first find the outliers in each class by a robust deep autoencoders-based outlier detector, through which not only did we get candidate mislabeled data but also a group of well-learned deep autoencoders. Then a reconstruction error minimization based approach is applied to these outliers to further filter and relabel the mislabeled data. The experimental results on MNIST dataset show that the proposed method could significantly reduce the false positive rate of outlier detection and improve the performance of both data cleaning and classification in the presence of label noise.**

***Index Terms*—label noise, outlier detection, robust deep autoencoder, reconstruction error minimization**

I. Introduction

A large-scale dataset with good annotation is the foundation of supervised learning algorithms. As collecting such a reliable dataset by expert manual labeling are often expensive and time-consuming, nowadays some simple and convenient alternative methods are commonly adopted, such as retrieving keywords from the web like *WebVision* dataset[1], or using crowd-sourcing platforms like Amazon Mechanical Turk[2]. However, these non-expert methods may result in a dataset with various degree of label noise. As a consequence, the classification performance of a system trained with such data may be deteriorated [1]. Moreover, due to the extra uncertainty caused by the label noise, the complexity of underlying models would generally be increased to account for such noise, which means that more samples are needed for effective learning [2].

This work is partially supported by National Science Foundation of China (61672280,61373060, 61732006), AI+ Project of NUAA (56XZA18009), Jiangsu 333 Project (BRA2017377), Qing Lan Project and 6140312020413.

[1]https://www.vision.ee.ethz.ch/webvision/

[2]https://www.mturk.com

To reduce the influence of label noise, cleaning the dataset is a natural way. Some methods clean the mislabeled data by directly using outlier detection techniques like Local Outlier Factor (LOF) [3] or One-Class Support Vector Machines (OC-SVM) [4] where they treat the data with noisy labels as outliers in their corresponding class. However, not all of the outliers are the mislabeled data. For example, according to the definition of outliers, '*An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism*' [5], a sample in a boundary region of a class may meet the definition but is not necessarily contains label noise.

Based on this idea, in this paper, instead of treating outliers as data points with label noise, we only consider them as *candidate* mislabeled data, and use another step to verify whether they are really with incorrect labels. To this end, we proposed a novel and effective method to reduce the effect of label noise, by combining outlier detection techniques with minimum reconstruction error. Specifically, we first partition the dataset based on their labels, and then use a state-of-the-art outlier detection method based on robust deep autoencoder (RDA) to detect the outliers in each subset, and treat them as candidate mislabeled data. Finally, we use a reconstruction error minimization based method to further verify such candidates by checking whether they can be reconstructed well by their corresponding class.

The remaining part of this paper is organized as follows. In Sect. II, the related work about solving label noise problem is briefly reviewed. The proposed pipeline of our method is introduced in Sect. III. In Sect. IV, we show and analyze various experimental results compared with the state-of-the-art methods. In Sect. V, we conclude the paper.

II. Related Work

In the literature of solving label noise problem, there are two main approaches, which are label noise cleaning methods and label noise robust methods respectively. Traditionally, cleaning dataset is a major and direct way for this. Those methods first try to find out data with noisy labels and then filter or relabel them to get a cleaner dataset. Some methods learn a complementary classifier and treat the misclassified data as label noise data [6], [7]. However, the misclassification based

cleaning methods sink into a *chicken-and-egg* dilemma where the misclassification analysis relies on a good classifier while a dataset with label noise may get a poor classifier. Another work is proposed in [8] where they introduce human expert to check the suspicious data from the support vectors of SVM, since the support vectors contain almost all of the mislabeled data. A further work tries to reduce the number of suspicious data to be checked based on active learning [9]. Although expert based methods work well in practice, it is impossible or hard to get human expert in some particular scenes. From this point of view, the proposed method can correct the mislabeled data automatically.

As for label noise robust methods, they try to make the mislabeled data less harmful to the model rather than cleaning them. Avoiding overfitting techniques are introduced to make the classifier not too sensitive for training set, such as robust loss functions [10] and ensemble learning [11] which can only suit for simple cases. On the other hand, some methods are particularly designed for label noise problem using various robust optimization methods [12], [13]. However, these are mainly supervised methods in which the effect of each data point on the postulated model is carefully controlled by design but at the risk of reduced learning efficiency.

III. The proposed method

A. Pipeline for Reducing Label Noise

The overall proposed pipeline for reducing label noise consists of two main steps:

- **Outlier detection with robust deep autoencoder (RDA)** For training data with label noise, we first partition these data based on their corresponding labels. Then a robust deep autoencoder is trained to detect outliers in each group. Finally, we obtain the outliers and the learned autoencoder for each group.
- **Reconstruction error minimization (REM) classifier** Based on the well-learned deep representation by autoencoders, a reconstruction error minimization (REM) based classifier is used to further get the label noise data from outliers while relabeling them.

B. Outlier detection with robust deep autoencoder (RDA)

The proposed pipeline is based on outlier detection techniques and most outlier detection methods (such as one class SVM [14], one class neural network [15] and etc.) can be naturally adopted. Here one of the state-of-the-art outlier detection methods, i.e., the robust deep autoencoder (RDA) [16], is chosen to detect the outliers. Suppose that we have a dataset X where each row represents an entry, and some of the entries are outliers. The main idea of RDA is to split X into two parts $X = L_D + S$, where L_D is the interpretable part that can be well reconstructed by deep autoencoder and S denotes the outliers which are difficult to reconstruct. The target function can be defined as follows:

$$\begin{aligned} &\min_{\theta,S} ||L_D - D_\theta(E_\theta(L_D))||_2 + \lambda||S^T||_{2,1} \\ &s.t. X - L_D - S = 0. \end{aligned} \tag{1}$$

where $E_\theta(\cdot)$ and $D_\theta(\cdot)$ denote the encoder and decoder of an deep autoencoder, $l_{2,1}$ norm for S^T is used as a row sparse regularized item, and λ is the balance factor. Moreover, we treat the j-th sample as an outlier if $||S(j,:)||_2$ equals to zero.

As for training the target function, back-propagation and proximal gradient are used to optimize $||L_D - D_\theta(E_\theta(L_D))||_2$ and $||S^T||_{2,1}$ respectively. To combine the two parts, Alternating Direction Method of Multipliers (ADMM) [17] and R. L. Dykstra's alternating projection method [18] are used. For a more specific algorithmic process can be seen in [16] which is not the focus of this paper.

C. Reconstruction error minimization (REM) classifier

After outlier detection by RDA in each class, not only the outliers in each class are picked out which are regarded as candidate label noise data, but also we get the K well-learned deep autoencoder $< E_{\theta_i}(\cdot), D_{\theta_i}(\cdot) >, i = 1, 2, \cdots, K$. Since the autoencoder can be seen as a template for the corresponding class which reflects the characteristics of each class to a certain extent, we utilize this kind of information to further filter the true label noise data from outliers.

Particularly, for an outlier x detected by RDA, we first predict its label y_{true} by simply calculating the reconstruction error on each deep autoencoder and then assign it to the class which has the minimum reconstruction error, as follows:

$$y_{true} = argmin_{i=1,2,3\ldots,K} \; ||D_{\theta_i}(E_{\theta_i}(x) - x||^2. \tag{2}$$

The reconstruction error is a good indictor to distinguish real mislabeled data from other outliers (e.g., data in the boundary region of its own class), as data points from a class tend to get a lower reconstruction error from its own class than from other classes.

Then, we determine whether an outlier is with label noise or not by an indicator function:

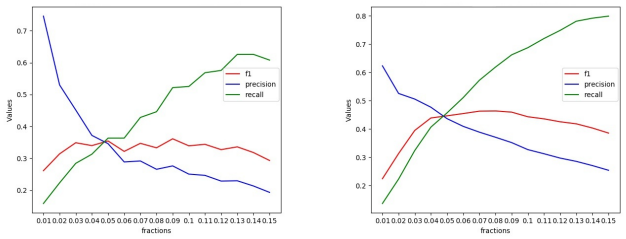
$$I(\mathbf{x}) = \begin{cases} 1 & y \neq y_{true} \\ 0 & otherwise \end{cases} \tag{3}$$

where y is their initial label and y_{true} is the predicted label which is regarded as the true label. An outlier will be regarded as a label noise data if y is not equal to y_{true}. Note that we can also regard the (2) as a classifier which can be used to relabel the mislabeled data and classify the test data.

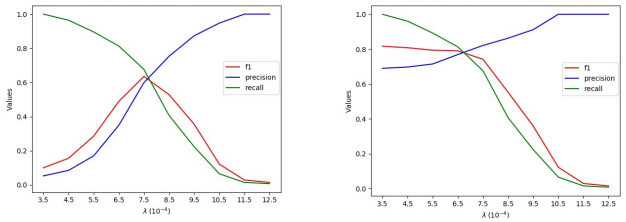
IV. Experiment and Results

A. Dataset and Evaluation Metrics

1) Dataset: In order to verify model performance, we use MNIST dataset to conduct our comparative experiments. MNIST dataset as a well-known digit dataset consists of 60000 training data and 10000 test data. Each image has 28 $\times$ 28 pixels. Since a training set with label noise is needed, we inject label noise following the protocol introduced in [1]: 1) randomly select instances per class and 2) flip the labels into one of the other remaining labels. Note that only the training set contains label noise, while the test set is clean.



(a) iForest (b) OC-SVM



(c) RDA (d) LN-RDA

Fig. 1. Performance of outlier detection for different models. The best F1-score for these models from (a) to (d) are 0.36, 0.46, 0.64, 0.82.

2) Evaluation Metrics: Hereby, we use accuracy, precision, recall, and F1-score as our evaluation metrics, which are defined as follows:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1\text{-}score = \frac{2 * TP}{2 * TP + FP + FN} \tag{7}$$

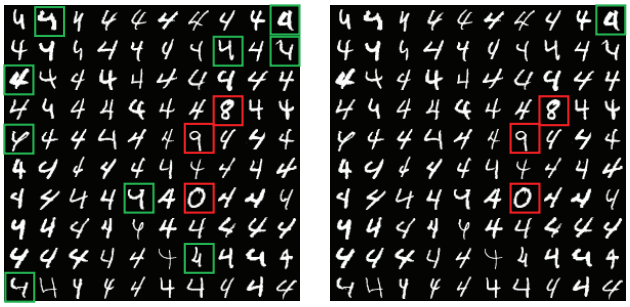
where TP, FP, TN, and FN are the true positive, false positive, true negetive, and false negetive respectively.

B. Label Noise Cleaning

Hereby, we demonstrate the effectiveness of our proposed method denoted as LN-RDA in label noise cleaning task. Firstly, performance of mislabeled data detection is compared with classical and state-of-the-art methods, which are Isolation forest (**iForest**), one class SVM (**OC-SVM**), and robust deep autoencoder (**RDA**).

Parameter settings for each model are shown as follows. For iForest, the number of trees is 100 and fractions is set from 0.01 to 0.15. As for OC-SVM, we select RBF kernel function for SVM and the same setting for fractions as iForest. Since our proposed LN-RDA is based on RDA method, they have the same setting for the model parameters of deep autoencoder and the λ we choose is from 0.00035 to 0.00125.

The comparison of experimental results are shown in Fig. 1 where 5% label noise is added. It can be seen that our proposed LN-RDA method gets the best F1-score among these methods.



(a) RDA (b) LN-RDA

Fig. 2. Label noise detection results for training data labeled '4'. All the outliers detected by RDA would be treated as data points with label noise, indicated by a square, where a red square indicates a true positive (i.e., a data point with label noise) while a green square indicates a false positive (i.e., a data point with correct label but are falsely identified as incorrect). One can see that generally a outlier detection method would lead to lots of false positives but our LN-RDA significantly reduces them.

Compared with the RDA method, LN-RDA has a higher precision which shows that the minimum reconstruction error criterion is effective for reducing the number of false positive instances. A more intuitive improvement can be seen in Fig. 2. As for the value of recall in LN-RDA, it is almost the same as RDA, since it is restricted to the effectiveness of outliers detection by RDA method.

Then we use the accuracy of relabeling the mislabeled data as the overall result. We relabel the mislabeled data by (2) as our cleaning result and compare it with several classical models which are ICCN-SMO [19], TC-SVM [8], and ALNR [9]. The hyper-parameters settings of these models are based on corresponding papers. Note that these methods learn from the annotation from human expert to relabel the mislabeled data, while in this respect, our method is completely automatic in the sense that we do not assume the existence of such supervised information.

TABLE I
PERFORMANCE FOR RELABELING MISLABELED DATA ON MNIST DATASET WITH 5% LABEL NOISE.

Methods	ICCN-SMO [19]	TC-SVM [8]	ALNR [9]	LN-RDA (ours)
Time cost (s)	938.6	352.1	70.8	**6.5**
Accuracy (%)	65.47	95.45	94.10	**98.06**

Table I gives the accuracy and time cost. It can be seen that although no human expert correction, our proposed method achieves the best accuracy compared with other human expert based methods. Moreover, since the relabeling process is conducted by human expert in these comparison methods, our model has the absolute advantage in time cost.

C. Classification in the Presence of Label Noise

In this section, we compare the classification performance for test set when the training set contains label noise. In the implementation of our proposed method, minimum reconstruction error classifier is also used as shown in (2). Note

that we select the optimal λ which gets the best F1-score in RDA for each class. Moreover, we train a normal deep autoencoder in each class and classify the test data by the proposed classifier. And we denote it as LN-DA which is used as the baseline method. Besides these, we also compare our method with two types of state-of-the-art methods, including cleaning based methods (TC-SVM [8] and ALNR [9]) and noise robust methods (L1-norm [20], BML [21], and RNCA [22]). In addition, the performance of all compared methods is based on the original implementation by corresponding papers and the related hyper-parameters are selected by cross-validation. We repeat thirty times for each experiment and report the mean and standard deviation of the accuracy.

TABLE II
CLASSIFICATION PERFORMANCE (%) ON MNIST DATASET WITH DIFFERENT LABEL NOISE.(THE ASTERISKS INDICATE A STATISTICALLY SIGNIFICANT DIFFERENCE BETWEEN THE SECOND BEST METHOD AND THE PROPOSED METHOD AT A SIGNIFICANCE LEVEL OF 0.05)

Methods	Noise level(%)				
	0	5	10	20	30
TC-SVM	98.27±0.06	95.83±0.06	94.95±0.16	91.45±0.20*	84.43±0.27*
ALNR	98.27±0.10	95.19±0.10	94.47±0.21	90.21±0.27	83.39±0.33
L1-norm	98.00±0.10	96.67±0.15	95.12±0.22*	89.78±0.31	81.39±0.49
BML	97.95±0.10	96.58±0.16	95.08±0.23	89.68±0.31	81.63±0.49
RNCA	98.15±0.11	96.73±0.17*	95.05±0.22	89.77±0.31	82.31±0.49
LN-DA	**99.12±0.15**	95.37±0.20	94.28±0.28	86.33±0.40	79.26±0.55
LN-RDA	99.04±0.11	**98.06±0.18**	**96.24±0.25**	**92.50±0.37**	**86.39±0.50**

Tabel 2 shows that our proposed method achieves the best performance consistently at both low-level and high-level noise compared with other methods, indicating the effectiveness of the pipeline in dealing with label noise problem. When the training set is clean, the baseline method LN-DA gets the best classification accuracy which shows the minimum reconstruction error based classifier is a suitable choice. However, due to the effect of label noise, the performance of LN-DA gets worse, especially in the high noise situation.

As for label noise cleaning methods, they achieve higher accuracy compared to the baseline method, as they learn the classifier using a cleaner training set. On the other hand, label noise robust methods achieve better results at low noise level compared with cleaning based methods. However, these methods perform worse at higher noise levels, highlighting the difficulty of obtaining reliable point estimation under the high-level label noise.

V. CONCLUSION

In this paper, based on a simple but novel idea that outlier detection can be regarded as a preliminary process for detecting candidate mislabeled data, we proposed to use minimum reconstruction error to further verify the truth of being contaminated of the detected data, as for a data point with correct label, the likelihood being reconstructed with minor error by a model of its own category would be much higher than a data point with incorrect label. Various experiments on the MNIST dataset show that our proposed method significantly reduce the false positive rate for an approach that naively applying an outlier detection algorithm to identifying the data points with label noise. We also show that our method outperforms several state-of-the-art methods in both data cleaning and classification, in the presence of label noise.

REFERENCES

[1] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2014.

[2] D. Wang and X. Tan, "Robust distance metric learning via bayesian inference," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1542–1553, 2018.

[3] H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, "Enhancing data analysis with noise removal," *IEEE Transactions on Knowledge & Data Engineering*, vol. 18, no. 3, pp. 304–319, 2006.

[4] H. Lukashevich, S. Nowak, and P. Dunker, "Using one-class svm outliers detection for verification of collaboratively tagged image training sets," in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 682–685.

[5] D. M. Hawkins, *Identification of outliers*. Springer, 1980, vol. 11.

[6] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Data cleaning for classification using misclassification analysis," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 14, no. 3, pp. 297–302, 2010.

[7] R. Pruengkarn, K. W. Wong, and C. C. Fung, "Data cleaning using complementary fuzzy support vector machine technique," in *International Conference on Neural Information Processing*. Springer, 2016, pp. 160–167.

[8] S. Fefilatyev, M. Shreve, K. Kramer, L. Hall, D. Goldgof, R. Kasturi, K. Daly, A. Remsen, and H. Bunke, "Label-noise reduction with support vector machines," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 3504–3508.

[9] R. Ekambaram, S. Fefilatyev, M. Shreve, K. Kramer, L. O. Hall, D. B. Goldgof, and R. Kasturi, "Active cleaning of label noise," *Pattern Recognition*, vol. 51, pp. 463–480, 2016.

[10] Y. Wu and Y. Liu, "Robust truncated hinge loss support vector machines," *Journal of the American Statistical Association*, vol. 102, no. 479, pp. 974–983, 2007.

[11] G. Rätsch, B. Schölkopf, A. J. Smola, S. Mika, T. Onoda, and K. R. Müller, "Robust ensemble learning for data mining," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2000, pp. 341–344.

[12] W. Zhang, D. Wang, and X. Tan, "Data cleaning and classification in the presence of label noise with class-specific autoencoder," in *International Symposium on Neural Networks*. Springer, 2018, pp. 256–264.

[13] D. Wang and X. Tan, "Bayesian neighborhood component analysis," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 7, pp. 3140–3151, 2018.

[14] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[15] R. Chalapathy, A. K. Menon, and S. Chawla, "Anomaly detection using one-class neural networks," 2018.

[16] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 665–674.

[17] Boyd, Vandenberghe, and Faybusovich, "Convex optimization," *IEEE Transactions on Automatic Control*, vol. 51, no. 11, pp. 1859–1859, 2006.

[18] J. P. Boyle and R. L. Dykstra, *A Method for Finding Projections onto the Intersection of Convex Sets in Hilbert Spaces*. Springer New York, 1986.

[19] U. D. Rebbapragada, "Strategic targeting of outliers for expert review," Ph.D. dissertation, Tufts University, 2010.

[20] H. Wang, F. Nie, and H. Huang, "Robust distance metric learning via simultaneous l1-norm minimization and maximization," in *International Conference on Machine Learning*, 2014, pp. 1836–1844.

[21] L. Yang, R. Jin, and R. Sukthankar, "Bayesian active distance metric learning," *arXiv preprint arXiv:1206.5283*, 2012.

[22] D. Wang and X. Tan, "Robust distance metric learning in the presence of label noise." in *AAAI*, 2014, pp. 1321–1327.