Label Noise Detection System Against Label Flipping Attack

Ning Cheng 1,2 · Hongpo Zhang 1,2 · Zhanbo Li 1,2

Received: date / Accepted: date

Abstract The label flipping attack is a special poisoning attack in the adversarial environment. By adding noise to the data, it destroys the learning process of the model and affects the decision-making performance of the model. Recent literature work uses semi-supervised learning techniques to defend against label flipping attacks. However, these methods require a clean dataset to achieve their goals. This study proposes a novel label noise processing framework to correct the labels of contaminated samples in the data preprocessing stage. Based on five real UCI datasets, we evaluated the effectiveness of the semi-supervised defensive label noise correction algorithm based on the AdaBoost algorithm (AdaSSL). With a noise ratio of 0~20%, we compared the classification performance of six classic machine learning algorithms (NB, LR, SVM, DT, KNN and MLP) under the AdaSSL defense algorithm. The results show that compared with the most advanced semi-supervised defense algorithms in the literature, our algorithm does not need to use an additional dataset. At a noise ratio of 10%, the AdaSSL algorithm is significantly better than state of the art label noise defense technology.

Keywords Class noise detection · Machine learning · AdaBoost algorithm · Semi-supervised

1 Introduction

Whether the training data is clean plays a role in the performance of the supervised machine learning model. Therefore, preprocessing the data before model training is a very critical step. In the adversarial environment, attackers use the vulnerability of uncertain or untrustworthy data to launch attacks on machine learning [24]. Poisoning attack is a kind of adversarial attack, which affects the learning of the model by tampering with the features and labels of the training data [2, 26]. Research against machine learning shows that by sacrificing a small part of training data, optimal poisoning attacks can effectively reduce the model (such as SVM [9], feature selection method [16], DNN [18]) classification performance [20]. The label flipping attack is a very harmful poisoning attack, which deceives the machine learning classifier by tampering with the labels of the training data. Biggio et al. [4] proposed an adversarial-rotating label flipping attack algorithm to destroy the learning process of SVM. Xiao et al. [27] increased the classification error of the classifier on the attack-free dataset by finding a set of optimal label flip sets. Taheri et al. [24] evaluated the impact of a label flipping attack algorithm based on Silhouette clustering on Android malware classification.

It is no general defense mechanism against poisoning attacks in security applications. The current research work is mostly based on a specific application to design a defense mechanism. For example, Paudice et al. [20] use outlier detection algorithms to effectively remove suspicious points in the data, but it is not suitable for dealing with adversarial label flipping attacks. Literature [3] uses principal component analysis and data dimensionality reduction to enhance the robustness of
machine learning, but this method is not suitable for large datasets.

The label flipping attack is a destructive label noise, which brings many potential negative effects on classification. For example, label noise will cause the accuracy of predicted data to decrease, increase the complexity of model learning, and so on. There are currently many label noise processing technologies. In [31], the author proposes an adaptive voting correction (AVNC) algorithm to detect noise label. Li et al. [15] proposed a new label noise correction technology—MCC. The author combined the advantages of K-Means clustering algorithm and KNN classification algorithm to estimate the noise rate in the data and divided it into three levels. The correction techniques used for noise at different levels are different. Compared with the previous noise correction technology, this algorithm is not only suitable for two classification problems, but also has good processing ability for the noise existing in multiple classifications.

Similar to most research areas of noise detection mechanisms, this article focuses on the label noise that exists in the binary classification task. From the perspective of label noise, we propose a label noise detection framework, mainly used to solve the degradation of classification performance caused by label flipping attacks. The goal of the label attack is to modify the sample labels of the training set to make the distribution deviate from the true distribution of the data. To solve this problem, we use the AdaBoost algorithm to assign greater weight to the noise samples to find such a type of noise samples, and then relabel these samples with the semi-supervised algorithm. Our contributions are as follows:

1. We propose a label noise detection framework, which reflects the main task of this research.
2. Aiming at the label flipping attack in the data, a semi-supervised label correction algorithm based on the AdaBoost algorithm is proposed. The effectiveness of the method is verified on five real UCI datasets.
3. Compared with the latest semi-supervised defense algorithm, the AdaSSL algorithm proposed in this paper does not need to use an additional clean dataset to correct noisy labels.

The rest of the paper is organized as follows. Section 2 introduces the related work of label noise defense. Section 3 elaborates the proposed label noise detection framework. In Section 4, we show the experimental performance of the algorithm on real datasets. Section 5 compares our algorithm with the latest defense algorithm. Section 6 summarizes the full text and draws our conclusions.

2 Related work

We review the related work of label noise defense from label noise detection algorithm and label flipping attack defense strategy.

2.1 Label noise detection algorithm

The data in practical applications is easily contaminated by label noise, which affects the training of model and reduces the classification accuracy of the prediction stage [9]. The current technology for processing label noise mainly includes three categories: 1) The inherent robustness of the learning algorithm; 2) Data cleaning method; 3) Label noise tolerance algorithm. Researchers often pay attention to the inherent robustness of the loss function or design a new robust loss function to deal with the negative effects of random label noise.

Reference [12] studied the tolerance of risk minimization to label noise and proved the robustness of the 0-1 loss function to uniform and nonuniform noise. The experimental results show that when the loss function has sufficiently large parameter values, the sigmoid loss, ramp loss and probit loss are also tolerant of nonuniform noise. In the multi-classification task, Aritra et al. [11] showed that the loss function based on the absolute value of the error is inherently robust to label noise in neural network learning. However, in the case of using MAE loss, training the network will become very slow, because the gradient will quickly saturate during training. Zhang et al. [32] proposed a new noise robust loss function—truncated Lp loss, which is an extension of the MAE and CCE loss functions, combining the robustness of MAE to noise and the fast convergence speed of the CCE function Features. The ensemble method improves the classification performance of the model by correcting the error of a single classifier. Sluban et al. [23] studied the relationship between integrated diversity and noise detection performance. Studies have shown that using majority voting schemes to increase diversity of the system will not bring better noise detection performance. In the noise detection system based on consensus voting, different systems can obtain higher noise detection recall rates and higher F1-score values. Samami et al. [21] improved the shortcomings of the above two voting schemes. They proposed a new classification noise detection strategy, a highly consistent voting filtering method with a hybrid strategy. Using five classification algorithms such as KNN, SVM, DT, NB and SVM to jointly label the data, the noise is divided into three categories: strong noise, semi-strong noise and weak noise.
The superiority of the proposed method is proved under two noise levels of 10% and 15%. Aiming at a large amount of unlabeled and noisy data in the data, Yan et al. [28] proposed an effective robust semi-supervised integrated learning method (ROSSL). First, they use a set of weak annotators to generate a large number of pseudo-labels for unlabeled data to approximate the real labels of the labeled data. Then they use multi-label kernel learning MLKL for weighted label aggregation, which has low time complexity. Aiming at uniformly distributed noise and normally distributed noise, Nicholson et al. [19] proposed a self-trained noise correction algorithm (STC) and a cluster-based noise correction algorithm (CC). Experiments were conducted on binary, multi-class, and crowdsourced datasets, and the results showed that CC showed better performance under the influence of different levels of noise. Li et al. [15] improved the noise correction algorithm based on clustering. They combined the advantages of supervised learning and unsupervised learning algorithms and proposed a new method to improve data quality. Using the KNN algorithm to estimate the noise in the data Level and combine this noise estimation with the correction process.

Studies have shown that smoothing technology can also effectively improve the influence of label noise on the model. Therefore, Lukasik et al. [17] proposed a method that can use label smoothing technology to deal with label noise. Moreover, the author proves through experiments that the label smoothing technology can improve the performance of the model under label noise, and it is quite competitive with the loss correction technology under label noise.

Existing noise detection and recognition techniques usually rely on supervision, but this supervision is not scalable. For this reason, Sharma et al. [22] proposed an unsupervised learning method using Markov random fields to detect label noise and verified its effectiveness through experiments. This method builds a dependency model to estimate the posterior probability of an instance being mislabeled in a given data set and ranks the instances according to the posterior probability.

2.2 Label flipping attack defense method

Unlike the label noise detection technology introduced in the previous section, data filtering and robust learning are two commonly used algorithms to defend against poisoning attacks for the purposeful destruction of data by attackers [6]. Data filtering is mainly to identify contaminated samples in the training set and remove them, while the purpose of robust learning is to improve the robustness of the model to noisy data. Because the data will change the geometric features of the data after the induced attack, reference [6] uses data complexity to describe the geometric characteristics of the data set. The author uses three different data complexity methods to describe the characteristics of the dataset from different angles and distinguishes contaminated samples from uncontaminated samples based on the data complexity measurement. The performance of this method is better than the current data filtering method and is suitable for the security field. For label flipping attacks, Paudice et al. [20] proposed a suspicious point data detection and marking algorithm to reduce the negative impact of the attack. The main idea of this algorithm is to use KNN to assign a label to each instance in the training set, but its performance is limited. Based on the characteristics of semi-supervised learning technology, Taheri et al. [24] proposed two semi-supervised algorithms to correct incorrect labels against label flipping attack.

3 Semi-supervised learning label correction framework based on AdaBoost

The following will introduce the semi-supervised learning label correction framework based on AdaBoost as shown in Figure 1. Our proposed method consists of two main modules. The first module uses the AdaBoost algorithm to calculate the weights of training samples containing noisy labels. It inputs a set of samples with larger sample weights to the semi-supervised learning module. Then in the second part, we use the advantages of semi-supervised learning algorithm to detect and relabel the labels of this group of samples. The technologies used in this framework are discussed in detail below.

3.1 Problem definition

We considered the label flipping attack in the binary classification task. Assume that the samples in the training dataset \( S_{tr} = \{(x_1, y_1'), \ldots, (x_i, y_i'), \ldots, (x_m, y_m')\} \) come from unknown distribution \( D \). Where \( m \) represents the number of training samples, \( x_i \) is the feature representation of the training samples, and \( y_i \in \{-1, +1\} \) is the sample label contaminated by the label. Let \( y_i \) be the true label corresponding to the sample \( x_i \). If the sample \( x_i \) is contaminated by label noise, then \( y_i' = y_i \); otherwise, \( y_i' = -y_i \).
3.2 classification algorithm

We used several classic supervised machine learning techniques \(^{25}\) to classify the binary samples in this study. Each of these algorithms has different characteristics, and we aim to use these algorithms to make a more comprehensive comparison. According to the advantages of the parameters in the learning algorithm, it can be divided into parametric classification and non-parametric classification. In this article, we use two parametric classification methods (i.e. LR and NB) and four non-parametric classification algorithms (SVM, DT, KNN and MLP).

To facilitate use and calculation, the naive Bayes classifier adds the assumption of conditional independence based on the Bayes classifier, which is widely used in text classification and other fields \(^{5}\). The imbalance between data categories is a big problem often encountered in machine learning classification. Logistic regression is a linear classifier that is suitable for large samples, and there is an imbalance between data \(^{29}\). SVM is a widely used supervised learning technique, which has been successfully applied in various fields \(^{7}\). KNN is robust to noisy data. It is an example-based learning algorithm that determines the category of test sample by comparing the labels of the K training samples that are closest to the test sample \(^{13}\). The working method of artificial neural network is similar to the way the human brain makes decisions. It uses a more compact network structure to improve the classification accuracy of the model \(^{25}\). Compared with neural networks, de-
cision trees are a highly understandable model. C4.5 and CART are two decision tree techniques that are currently widely used.

3.3 Defense strategy

In this section, we discuss the label noise detection framework against label flipping attack. Firstly, capture a labeled set of noise samples based on sensitivity of the AdaBoost algorithm to noise data; and then use the semi-supervised learning algorithm to relabel the label of the sample set. The label noise detection method is shown in Algorithm 1.

Algorithm 1 AdaSSL

Input: $X_{train}, Y_{noise}, \text{weak classifier } C, \text{iterations } T$
Output: Corrected_label $Y$

1: Initialize the sample weight of each training sample $\omega_i = \frac{1}{m}$
2: for $i = 1$ to $T$ do
3: Use the weak classification algorithm $C$ to train the model to get the classification error rate $\epsilon_t = \sum_{i=1}^{m} \omega_i I(C_t(x_i) \neq y_i)$
4: Use the AdaBoost algorithm to iteratively update the weights of the incorrectly classified samples of classifier $C: W^{t+1} = \left\{ \begin{array}{ll}
\omega_i e^{-\epsilon_t} & C_t(x_i) = y_i \\
\frac{\omega_i e^{-\epsilon_t}}{Z_t} & C_t(x_i) \neq y_i
\end{array} \right.$
5: end for
6: After $T$ iterations, the final classifier is obtained $C(x) = \text{sign} \left( \sum_{t=1}^{T} \alpha_t C_t(x) \right)$
7: Find a set of samples with larger weights through the final classifier $U$
8: Use semi-supervised learning algorithm to relabel $U$ to get the label set: Corrected_comment $Y$

3.3.1 AdaBoost Algorithm

The AdaBoost algorithm is a Boosting ensemble method for two classifications, first proposed by Freund et al. The core of the algorithm is to train multiple weak classifiers based on the same training set, and linearly combine these weak classifiers into a strong classifier through $T$ iterative update until a sufficiently small classification error rate is achieved. The error rate of each base classifier is determined by the following Eq. 1:

$$
\epsilon_t = \sum_{i=1}^{m} \omega_i I(C_t(x_i) \neq y_i)
$$

where $\omega_i$ is the weight corresponding to the training sample, and $C_t(x_i)$ is the base classifier obtained after $t$ iterations in the training set.

The advantage of the AdaBoost algorithm is that it can iteratively update the sample weights of each round. For the samples classified incorrectly in the previous round, AdaBoost will give them a larger sample weight, and the correct samples will reduce their weight. The update process of training sample weights is shown in Eq. 2:

$$
W^{t+1} = \left\{ \begin{array}{ll}
\omega_i e^{-\epsilon_t} & C_t(x_i) = y_i \\
\frac{\omega_i e^{-\epsilon_t}}{Z_t} & C_t(x_i) \neq y_i
\end{array} \right.
$$

where $\alpha_t$ is the coefficient of the base classifier $C_t(x_i)$, which represents the importance of the base classifier to decision-making; $Z_t = \sum_{i=1}^{m} \omega_i e^{-\alpha_t y_i C_t(x_i)}$ is the normalization factor for sample weight.

The AdaBoost algorithm is very sensitive to noisy data and abnormal data. During the iterative update process of model training, AdaBoost will adjust the sample's weight according to the previous training result. Therefore, given the label noise in the training dataset, we use the iterative principle of AdaBoost algorithm to find a set of suspicious samples with larger sample weights. Next, we input these noise data into the semi-supervised learning module and use the advantages of semi-supervised learning technology to further process sample labels.

3.3.2 Semi-supervised learning method

Semi-supervised algorithms are between supervised learning and unsupervised learning. Generally speaking, the tasks of semi-supervised learning are the same as those of supervised learning, and the tasks contain clear goals, such as classification. The semi-supervised learning algorithm is suitable for classification tasks that contain a small number of labeled samples, and the remaining samples are unlabeled. The marked sample can be expressed as $S_{label} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\}$, The unlabeled sample set are $S_{unlabel} = \{x_{l+1}, x_{l+2}, \cdots, x_{l+t}\}$. Train the unlabeled sample set by using the labeled sample set, and finally get a classifier to classify the test sample.

The use of semi-supervised learning techniques to label unlabeled samples is generally based on two assumptions: clustering assumptions and manifold assumptions. The clustering hypothesis refers to dividing categories according to the distance between samples; the core idea of the manifold hypothesis is that the data samples in the same local domain have similar characteristics, so the categories to which these samples belong should also be similar. At present, the more popular semi-supervised learning methods include self-training methods, semi-supervised SVM methods, generative model methods, graph-based semi-supervised learning methods, etc.
The semi-supervised learning technique used in our label noise detection framework is based on the manifold hypothesis, using the labeled dataset to train the model to correct label on the noise data.

4 Experiment

To evaluate the effectiveness of the proposed label noise detection framework against label flipping attack, we conduct experiments on five standard datasets in the UCI machine learning library. First, the entropy method based label flipping attack [30] is used to artificially introduce 5%, 10%, 15% and 20% label noise; Then we use the machine learning algorithm in section 3.2 to classify the dataset and compare their performance under label attack and defense method; Next, we use the label noise detection framework proposed in Section 3 to correct the noisy dataset; Finally, we again use the above-mentioned machine learning techniques to classify the dataset. The effectiveness of our method is verified by comparing the classification performance before and after the classifier.

4.1 Dataset description

We conducted experiments on five UCI datasets: Spambase, Breast-w, Kr-vs-kp, Diabetes and Biodeg. These datasets have different feature distributions, different number of instances and feature types. Table 1 describes the details of these datasets. We divide the training set and test set according to a 4:1 ratio. Note that the dataset used only standardizes the features of data, and there is no other further processing.

| Datasets   | Instances | Positive | Negative | Features | Features.type |
|------------|-----------|----------|----------|----------|---------------|
| Spambase   | 4,601     | 1,813    | 2,788    | 57       | numeric       |
| kr-vs-kp   | 3,196     | 1,527    | 1,669    | 36       | nominal       |
| Diabetes   | 768       | 268      | 500      | 8        | numeric       |
| Breast-w   | 699       | 241      | 458      | 9        | numeric       |
| Biodeg     | 1,055     | 356      | 699      | 41       | numeric       |

4.2 Evaluation metric

To visually show effectiveness of the proposed method, we use evaluation metrics such as Accuracy, F1-score, AUC, Recall, and Precision to show the classification performance of machine learning algorithms on these datasets. The following formulas (3)-(6) give the calculation formulas of these indicators. Among them, TP represents the number of positive samples correctly classified, TN represents the number of correctly classified negative samples, FP represents misclassification of negative samples as positive samples, and FN represents misclassification of positive samples as negative samples. Accuracy indicates the proportion of samples that are predicted to be correct to the total number of samples; For the original sample, Recall represents the probability that the actual positive sample is predicted to be a positive sample; Precision is an evaluation index for the prediction results, which represents the probability that the actual positive samples account for all the predicted samples as positive; F1-score is a comprehensive evaluation metric of recall and precision.

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (4)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (5)
\]

\[
F_1 - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)
\]

4.3 Experimental results

We evaluate the defense performance of the AdaSSL algorithm against label noise attack. We focused on the label flipping attack based on the entropy method. Among them, the noise ratio is defined as the percentage of contaminated samples in training samples.

Based on the UCI datasets used in this study, Figure 2 shows the accuracy results of classification model under the entropy method attack and label noise defense strategy. It can be seen from Figure 2 that in the attack interval of [0.0, 0.2], the accuracy value of the machine learning algorithm on the five datasets decreases in the range of 15% to 30%, which illustrates the effectiveness of based on entropy method attack method. Except for the NB classifier, the test accuracy of supervised classifier dropped by about 20% when the attack rate was 20%. It also shows that traditional machine learning classification methods are very sensitive to label noise attack. Figure 2 also shows that using the
semi-supervised learning label correction method based on AdaBoost proposed in this paper effectively reduces the attack effect. The performance of the classification model trained on the training set after data label correction is generally better than the classification algorithm under the label noise attack. However, on the Biodeg dataset, our method performs poorly when using the NB classifier. It is because the NB algorithm has better robustness to noise, and NB is a generative model. Using this attack method makes the distribution of the training data of the dataset closer to the model. Using this attack method makes the distribution of the training data of the dataset closer to the model. Using this attack method makes the distribution of the training data of the dataset closer to the model.

Figures 3 and 4 show the changes of $F_1$-score and AUC values on all UCI datasets under label noise pollution and defense strategies. $F_1$-score measurement is a comprehensive evaluation metric of recall rate and precision rate. It can be seen from figure 3 that for the drop in $F_1$-score value caused by label contamination attack, the AdaSSL label correction method proposed in this study significantly improved the label quality of contaminated samples. For example, on the Biodeg dataset, the $F_1$-score of the SVM model drops to 0 at a noise ratio of 20%. It is because the positive samples of test set are misclassified by the model due to influence of the entropy label noise. Through our proposed AdaSSL method to improve the label of the training data, from the experimental results, the $F_1$-score of the dataset has been significantly improved. The AUC value can be used to measure the robustness of the classification model to noise. It can be seen from Figure 4 that on these two class datasets, our method significantly improves the label quality of the data attacked by label contamination. Based on this research, the AdaSSL algorithm disinfects tainted data label, which improves the robustness of machine learning algorithms to noise data.

5 Discussion

To further demonstrate the effectiveness of AdaSSL defense algorithm, we will compare it with the latest three latest semi-supervised label attack defense algorithms (KSSD [20], LSD [24] and CSD [24]). We compared the performance of these defense methods under the entropy method based label flipping attack on the SpamBase dataset. We choose SVM, LR and MLP as three representative classification algorithms. As shown in Table 2 and Table 3, we respectively show the Accuracy and $F_1$-score values of these defense algorithms at 10% and 20% noise ratios. It can be found from the table that at an attack rate of 10%, the AdaSSL method proposed in this study has the best performance, followed by CSD, KSSD and LSD algorithms. The KSSD has the best defense performance at a noise ratio of 20%, followed by our AdaSSL algorithm. We also found that the performance of KSSD algorithm is the same regardless of the noise ratio of 10% or 20%. It is because the KSSD algorithm uses another clean validation dataset to correct noisy labels.

| Defense Algorithm | SVM | LR | MLP |
|-------------------|-----|----|-----|
|                   | Accuracy | $F_1$-score | Accuracy | $F_1$-score | Accuracy | $F_1$-score |
| LFA               | 85.12  | 78.08 | 83.71 | 75.96 | 85.78 | 79.50 |
| KSSD              | 86.21  | 80.13 | 84.58 | 77.81 | 86.32 | 80.31 |
| LSD               | 85.02  | 78.44 | 82.08 | 74.42 | 75.24 | 75.24 |
| CSD               | 87.40  | 82.11 | 86.10 | 80.31 | 89.03 | 84.72 |
| AdaSSL            | 89.36  | 84.43 | 88.82 | 80.02 | 89.79 | 87.33 |

| Defense Algorithm | SVM | LR | MLP |
|-------------------|-----|----|-----|
|                   | Accuracy | $F_1$-score | Accuracy | $F_1$-score | Accuracy | $F_1$-score |
| LFA               | 70.79  | 43.61 | 72.96 | 51.08 | 76.76 | 60.81 |
| KSSD              | 86.21  | 80.13 | 84.58 | 77.81 | 86.32 | 80.31 |
| LSD               | 88.40  | 86.00 | 82.91 | 44.94 | 86.56 | 51.72 |
| CSD               | 76.66  | 59.81 | 75.73 | 62.82 | 80.87 | 69.29 |
| AdaSSL            | 84.69  | 78.67 | 83.82 | 77.66 | 85.23 | 79.94 |

6 Conclusion

In this article, we design a novel noise detection framework to defend against label flipping attack. As the core component of the detection framework, the AdaSSL algorithm combines the sensitivity of the AdaBoost algorithm to noise and the features of the semi-supervised learning algorithm to classify unlabeled data to correct noisy labels. We verified the effectiveness of this method on five real UCI data sets, and compared the latest three semi-supervised defense algorithms. In the future, we hope to apply the researched detection framework to specific research fields.

Acknowledgements This research was partly supported by the Integration of Cloud Computing and Big Integration of Cloud Computing and Big Data, Innovation of Science and Education (grant no. 2017A11017), the Key Research, Development, and Dissemination Program of Henan Province (Science and Technology for the People) (grant no. 182207310002), and the Key Science and Technology Project of Xinjiang Production and Construction Corps (grant no. 2018AB017).

Author contributions: Hongpo Zhang contributed to the conception of the study; Ning Cheng performed the experiment;
Fig. 2 Accuracy results on UCI datasets
Fig. 3 $F_1$-score results on UCI datasets
Fig. 4 AUC results on UCI datasets
Hongpo Zhang contributed significantly to analysis and manuscript preparation; Ning Cheng performed the data analyses and wrote the manuscript; Zhanbo Li helped perform the analysis with constructive discussions.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Altinel, B., Ganiz, M.C.: A new hybrid semi-supervised algorithm for text classification with class-based semantics. Knowledge-Based Systems 108, 50–64 (2016)
2. Barreno, M., Nelson, B., Sears, R., Joseph, A.D., Tygar, J.D.: Can machine learning be secure? In: Proceedings of the 2006 ACM Symposium on Information, computer and communications security, pp. 16–25 (2006)
3. Bhagoji, A.N., Cullina, D., Mittal, P.: Dimensionality reduction as a defense against evasion attacks on machine learning classifiers. arXiv preprint arXiv:1704.02654 (2017)
4. Biggio, B., Nelson, B., Laskov, P.: Support vector machines under adversarial label noise. In: Asian conference on machine learning, pp. 97–112 (2011)
5. Biggio, B., Nelson, B., Laskov, P.: Poisoning attacks against support vector machines. arXiv preprint arXiv:1206.6389 (2012)
6. Chan, P.P., He, Z.M., Li, H., Hsu, C.C.: Data sanitization against adversarial label contamination based on data complexity. International Journal of Machine Learning and Cybernetics 9(6), 1039–1052 (2018)
7. Demidova, L., Klyueva, I., Sokolova, Y., Stepanov, N., Tyart, N.: Intellectual approaches to improvement of the classification decisions quality on the base of the svm classifier. Procedia Computer Science 103, 222–230 (2017)
8. Diab, D.M., El Hindi, K.M.: Using differential evolution for fine tuning naive bayesian classifiers and its application for text classification. Applied Soft Computing 54, 183–199 (2017)
9. Frénay, B., Verleysen, M.: Classification in the presence of label noise: a survey. IEEE transactions on neural networks and learning systems 25(5), 845–860 (2015)
10. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences 55(1), 119–139 (1997)
11. Ghosh, A., Kumar, H., Sastry, P.: Robust loss functions under label noise for deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
12. Ghosh, A., Manwani, N., Sastry, P.: Making risk minimization tolerant to label noise. Neurocomputing 160, 93–107 (2015)
13. Gupta, V., et al.: Recent trends in text classification techniques. International Journal of Computer Applications 35(6), 45–51 (2011)
14. Kotsiantis, S.B.: Decision trees: a recent overview. Artificial Intelligence Review 39(4), 261–283 (2013)
15. Li, B., Gao, Q.: Improving data quality with label noise correction. Intelligent Data Analysis 23(4), 737–757 (2019)
16. Liu, H., Ditzler, G.: Data poisoning attacks against mrrm. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2517–2521. IEEE (2019)
17. Lukasik, M., Bhojanapalli, S., Menon, A.K., Kumar, S.: Does label smoothing mitigate label noise? arXiv preprint arXiv:2003.02819 (2020)
18. Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wonggrassamee, V., Lupu, E.C., Rolli, F.: Towards poisoning of deep learning algorithms with back-gradient optimization. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 27–38 (2017)
19. Nicholson, B., Sheng, Y.S., Zhang, J.: Label noise correction and application in crowdsourcing. Expert Systems with Applications 66, 149–162 (2016)
20. Paudice, A., Muñoz-Gonzalez, L., Lupu, E.C.: Label sanitization against label flipping poisoning attacks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 5–15. Springer (2018)
21. Samani, M., Akbari, E., Abdar, M., Pliaiak, P., Nematzadeh, H., Basiri, M.E., Makarenkov, V.: A mixed solution-based high agreement filtering method for class noise detection in binary classification. Physica A: Statistical Mechanics and its Applications p. 124219 (2020)
22. Sharma, K., Donmez, P., Luo, E., Liu, Y., Yalınz, I.Z.: Noiserank: Unsupervised label noise reduction with dependence models. arXiv preprint arXiv:2003.06729 (2020)
23. Sluban, B., Lavrač, N.: Relating ensemble diversity and performance: A study in class noise detection. Neurocomputing 160, 120–131 (2015)
24. Taheri, R., Javidan, R., Shojafar, M., Pooranian, Z., Miri, A., Conti, M.: On defending against label flipping attacks on malware detection systems. Neural Computing and Applications pp. 1–20 (2020)
25. Thangaraj, M., Sivakami, M.: Text classification techniques: A literature review. Interdisciplinary Journal of Information, Knowledge & Management 13 (2018)
26. Xiao, H., Biggio, B., Brown, G., Fumera, G., Eckert, C., Rolli, F.: Is feature selection secure against training data poisoning? In: International Conference on Machine Learning, pp. 1689–1698 (2015)
27. Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C., Rolli, F.: Support vector machines under adversarial label contamination. Neurocomputing 160, 53–62 (2015)
28. Yan, Y., Xu, Z., Tsang, I., Long, G., Yang, Y.: Robust semi-supervised learning through label aggregation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30 (2016)
29. Yen, S.J., Lee, Y.S., Ying, J.C., Wu, Y.C.: A logistic regression-based smoothing method for chinese text categorization. Expert Systems with Applications 38(9), 11581–11590 (2011)
30. Zhang, H., Cheng, N., Zhang, Y., Li, Z.: Label flipping attacks against naive bayes on spam filtering systems. Applied Intelligence pp. 1–12
31. Zhang, J., Sheng, V.S., Li, T., Wu, X.: Improving crowdsourced label quality using noise correction. IEEE transactions on neural networks and learning systems 29(5), 1675–1688 (2017)
32. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels.
Advances in neural information processing systems 31, 8778–8788 (2018)