Machine Learning Model Data Poisoning Attacks and Countermeasures Research

Ruoheng xi

School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China

Abstract. Currently, machine learning models are widely used in natural language processing, image and video processing, and other fields. However, security issues such as data poisoning attacks threaten their performance and reliability. It is of great significance to study the defense mechanism. Therefore, this paper proposes a defense method based on dynamic network structure adjustment and adaptive learning weights and analyzes its effect in improving the robustness of model poisoning attacks. Experiments show that the accuracy of the model under poisoning attack drops from 100% to 76.67%, and the accuracy increases to 83.33% after adopting the defense mechanism. By dynamically adjusting the network structure and learning weights, the model can adapt to changes in data distribution, reduce the interference of poisoning data, and stabilize the decision boundary without obvious underfitting. This study provides an effective reference for improving the robustness of the machine learning model under data poisoning attacks, which helps to ensure the stable operation of the model in complex environments.

1 Introduction

With the wide application of machine learning technology in key fields such as healthcare, transportation, and finance, its security problem has gradually become a research hotspot. The training and deployment of machine learning models are highly dependent on data, however, data may face various security threats in the process of collection, storage and use, among which data poisoning attacks are particularly prominent. Data poisoning attacks interfere with the model training process by maliciously tampering with the training data, thus reducing the performance and reliability of the model, and may even trigger risks such as data leakage. Therefore, it is of great theoretical and practical significance to study how to effectively defend against data poisoning attacks and guarantee the robustness and stability of machine learning models[1].

Currently, the defense methods for data poisoning attacks mainly focus on data preprocessing, model robustness optimization and anomaly detection. For example, Ma et al. [2] studied data poisoning attacks and their defense methods in differential privacy learning, proposed a game model between attackers and defenders, and improved the robustness of the model by optimizing the defense strategy.[3] Müller et al. conducted a study on data poisoning attacks in regression learning, and proposed a defense mechanism

22122827@shu.edu.cn

based on statistical detection, which is able to identify and filter out some of the malicious data[4]. In addition, Chen et al. proposed an attack-independent defense method (De-Pois) to reduce the impact of poisoned data by optimizing the training process of the model[5]. However, these methods still have some limitations in practical applications, for example, data cleaning techniques make it difficult to completely identify malicious samples hidden in normal data, while model robustness optimization methods may increase model complexity and computational cost. Therefore, designing an efficient and reliable defense mechanism under the premise of guaranteeing model performance has become a key issue to be solved.

The purpose of this paper is to propose a defense mechanism based on dynamic network structure adjustment and adaptive learning weights to improve the robustness of machine learning models under data poisoning attacks. This paper first introduces the knowledge of data poisoning attacks, including their modeling approaches and attack targets. Then, the design ideas and implementation methods of the proposed defense mechanism are elaborated in detail. Through experimental validation on the Iris dataset, the effectiveness of the defense mechanism in reducing the impact of poisoning attacks on model performance is demonstrated. Finally, the main results of the study are summarized and the future research directions are outlined.

2 Data and methods

2.1 Data sources

The dataset used in this paper is the classic Iris (iris) dataset (https://gitcode.com/open-source-toolkit/9a64d/). The Iris dataset is one of the widely used benchmark datasets in the field of machine learning, and its dataset contains 150 samples, each of which represents measurements of an iris flower The Iris dataset is one of the most widely used benchmark datasets in machine learning. Each sample contains 4 features, Sepal Length, Sepal Width, Petal Length and Petal Width. In addition, each sample was accompanied by a label indicating the iris category to which the sample belonged (Setosa, Versicolour, Virginica). The dataset contains 3 categories with 50 samples in each category, which is a relatively even distribution and suitable for research on classification tasks.

Before the experiment, to improve the robustness and training efficiency of the model, abnormal data detection and removal are first performed: the k-NN algorithm is utilized to detect the data and remove possible abnormal data points. This method can effectively identify the samples with large differences from normal data, thus reducing the interference of potential noise on model training.

The data is then normalized: the feature values of the remaining data are normalized, and all the feature values are scaled to the interval [0, 1]. The normalization process can reduce the differences in magnitude between different features, improve the stability of the data, and help to speed up the convergence of the model.

2.2 Methods

2.2.1 Experimental Methods

In this paper, the research designs an adaptive defense mechanism based on a dynamic network structure, which dynamically adjusts the network structure during the training process in order to reduce the sensitivity of the model to the poisoning data. [6]Specifically, the model adaptively adjusts the structural parameters such as the number of layers and

neurons of the network during the training process according to the quality of the data and the performance of the model. For detected poisoning data, the model will automatically reduce its learning weights on these data, thus reducing the impact of poisoning data on model performance. This method not only improves the robustness of the model but also maintains the classification performance of the model.

The MLP model used in this paper contains one input layer, two hidden layers and one output layer. The input layer has 4 neurons corresponding to 4 features; the hidden layer has a size of 10 and uses the ReLU activation function; the output layer has 3 neurons corresponding to 3 categories:

2.2.2 Dynamic network structure

The number of hidden layer neurons in the initial network structure is 10. During the training process, every 20 epochs is an evaluation cycle, and the model performance is judged by calculating the accuracy of the current model on the training set. If the accuracy rate is lower than 0.85, the network structure will automatically adjust to halve the number of neurons in the hidden layer to reduce the complexity of the model; conversely, if the accuracy rate is at or above 0.85, the network structure will increase the number of neurons in the hidden layer to enhance the expressive power of the model[7].

2.2.3Adaptive defense mechanisms

By dynamically adjusting the network structure during the training process, the model can better adapt to the impact brought by data poisoning attacks. For example, when the model detects a drop in performance, the network size is reduced in time to reduce the interference of noisy data; and when the performance is stable and high, the network size is moderately increased to further enhance the model performance (Fig.1). This adaptive defense mechanism can effectively enhance the model's ability to resist poisoning attacks without relying on a large amount of additional data or computational resources, ensuring the stability and accuracy of the model in complex environments[8].

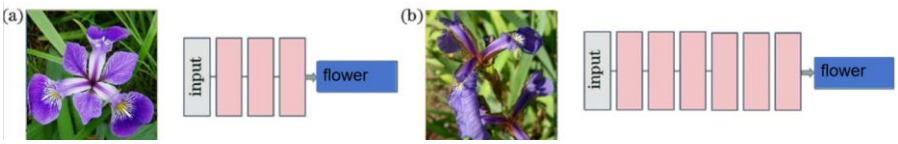

Fig. 1 Deep adaptive neural network that automatically adjusts the depth of inference according to the complexity of the input. (a) Network structure for processing simple inputs; (b) Network structure for processing complex inputs (Photo credit : Original)

2.2.4 Experimental procedure

The dataset is divided into a training set (20%) and a test set (80%) to train the model, and based on the defense mechanism of dynamic network structure adjustment and adaptive learning weights, the initial model, the model after data poisoning, and the model optimized by dynamic network structure adjustment and adaptive learning weights are tested and compared in terms of accuracy, respectively.

2.3 Assessment of Indicators

In this paper, classification accuracy as well as model complexity are used as evaluation metrics. The classification accuracy rate reflects the overall performance of the model on the test set, and the higher the value, the stronger the model's ability to classify data. In the scenario of a data poisoning attack, the change of the accuracy rate can intuitively reflect the impact of the attack on the model performance, as well as the effect of the defense mechanism on the enhancement of the model robustness[9].

Model complexity reflects the scalability and efficiency of the model in practical applications. When designing the defense mechanism, it is necessary to find a balance between improving robustness and maintaining the efficiency of the model, so as to avoid excessive model complexity due to the introduction of the defense mechanism, thus affecting its application value in practical scenarios[10].

3 Analysis of results

3.1 Analysis

Table 1 Accuracy results for each model

	test set	target	Adaptive defense
ACC	100%	76.67%	83.33

As shown in Table 1, on the normal test set, the classification accuracy of the model is 100%, which indicates that the model performs well without attacks, and the 100% accuracy may be because the dataset is very simple or the model still has an overfitting problem. Under poisoning attack, the classification accuracy of the model decreases to 76.67%, indicating that a poisoning attack has a significant effect on the model performance. After using the adaptive defense mechanism based on dynamic network structure, the classification accuracy of the model under poisoning attack increases to 83.33%. This indicates that the defense mechanism can effectively reduce the impact of poisoning attacks on the model performance and improve the robustness of the model.[1]

Figure 2 shows the data distribution of the Iris dataset before and after the poisoning attack. Specifically, Fig. 2(a) shows the distribution of the original training data, while Fig. 2(b) demonstrates the data distribution after the poisoning.

From Fig. 2(a), it can be seen that the distribution of the original training data is more centralized, and although there is overlap in the data points of different categories, there are no obvious anomalies on the whole. This indicates that in the absence of poisoning attacks, the distribution of the dataset is relatively uniform and the model can learn the features of the data better.

However, in Fig. 2(b), after the poisoning attack, the data distribution changes significantly. The data points become more dispersed, and anomalies appear in all categories of data.

In Fig. 2(a), the purple data points are mainly concentrated in the negative region of PCA Component 2, while in Fig. 2(b), these data points not only become more dispersed in the original region, but also some anomalies appear, which are distributed in the positive region of PCA Component 2, far away from the the original clustered region. The green data points are mainly concentrated in the positive value region of PCA Component 1 and the negative value region of PCA Component 2 in Fig. 2(a). In Fig. 2(b), these data points become more dispersed and some anomalies appear, which are distributed in the negative

value region of PCA Component 1 and the positive value region of PCA Component 2, far away from the original clustered region. The yellow data points are mainly concentrated in the positive value region of PCA Component 1 and the positive value region of PCA Component 2 in Fig. 2(a). In Fig. 2(b), these data points become more dispersed and some anomalies appear, which are distributed in the negative value region of PCA Component 1 and the negative value region of PCA Component 2, far away from the original clustered region.

These anomalies may be malicious data introduced by the poisoning attack, and their presence makes the data distribution more complex and increases the difficulty of model classification. This change in data distribution directly affects the accuracy of the model and leads to performance degradation.

By dynamically adjusting the network structure, the robustness of the model can be improved so that it can better adapt to changes in data distribution when facing poisoning data. Specifically, the model can dynamically adjust the number of neurons in the hidden layer according to the distribution of data, thus optimizing the model structure to adapt to the new data distribution. In addition, by dynamically adjusting the learning weights, the learning weights on the poisoning data can be reduced, thus reducing its interference with the model training[8].

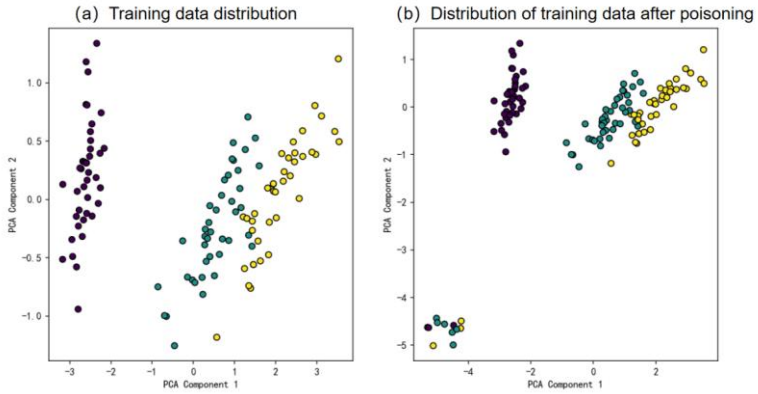

Fig. 2 (a) Distribution of training data; (b) Distribution of training data after poisoning (Photo credit : Original)

Figure 3 shows the decision boundaries of the model optimized on the training data. By PCA downscaling to a two-dimensional space, it can be visualized how the model classifies different categories of data. It can be seen that the decision boundary is relatively smooth and stable, with no obvious offset, overfitting, or underfitting phenomenon, indicating that the model is robust to poisoning data.

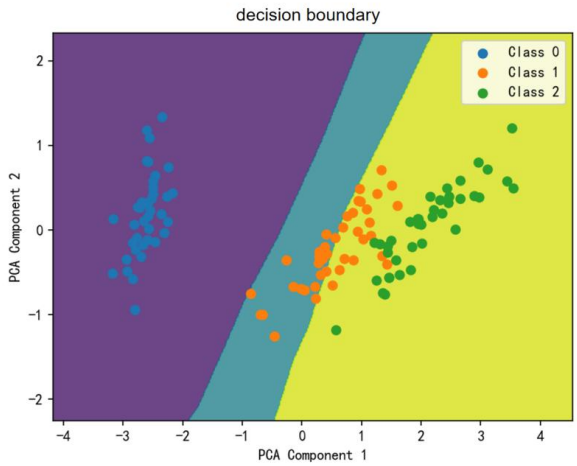

Fig. 3 Decision boundary after dynamic network restructuring and adaptive learning weight optimization (Photo credit : Original)

By dynamically adjusting the learning weights, the model can reduce the learning weights on the poisoning data, thus reducing the interference of the poisoning data on the model training. By dynamically adjusting the network structure (e.g., the number of neurons), the model is able to better adapt to changes in the data and reduce the impact of poisoning data on the model performance. As shown in Fig. 4,at each epoch or every few epochs, the model's performance (e.g., accuracy) is evaluated. If the model's accuracy on the training set is below a certain threshold (85), the number of neurons is reduced; if it is above that threshold, the number of neurons is increased.

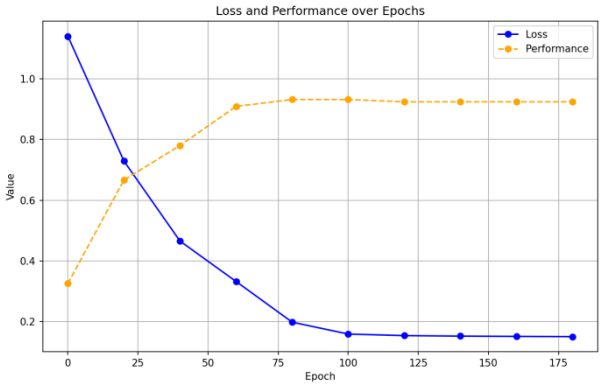

Fig. 4 Changes in loss value and accuracy with epoch continuously changing (Photo credit : Original)

3.2 Deliberation

The research strength of this paper is that it proposes a defense mechanism that combines dynamic network structure adjustment and adaptive learning weights to deal with data poisoning attacks. The mechanism improves the robustness of the model by dynamically adjusting the network structure and learning weights so that it can better adapt to changes in data distribution in the face of poisoned data.[8] Furthermore, by eliminating outliers or expanding the existing dataset to mitigate data poisoning, this active approach intentionally introduces controlled changes into the training dataset, such as random rotations,

translations, and noise additions. Data augmentation has the advantage of expanding the dataset while maintaining its inherent characteristics. In addition, the increase in sample size undermines any attempt to poison the dataset.

However, there are limitations, and the experimental results may be affected by factors such as the size of the dataset, the proportion of poisoning data, and the structure of the model. In practical applications, it is necessary to further verify the effectiveness of the defense mechanism in different datasets and attack scenarios. For example, medical models are particularly vulnerable to "data poisoning" attacks, mainly due to the strong concealment, low attack threshold, huge impact, unpredictable proliferation, and the limitations of existing testing methods.

In addition, current benchmark tests for healthcare AI are mainly based on multiple-choice questions and answers to frequently asked questions, but these tests are unable to detect the model's ability to generate harmful suggestions. Therefore, there is a need to further explore more robust defense mechanisms and develop detection and defense methods specifically for Text-to-SQL model backdoor attacks, including improved static analysis tools, dynamic behavior monitoring, and model sanitization techniques.

4 Conclusions

In this paper, the research investigates how to improve the robustness of machine learning models in the face of data poisoning attacks by proposing a defense mechanism that combines dynamic network structure adjustment and adaptive learning weights. In the study, this paper first uses the k-NN algorithm to initially detect the training data to remove obvious abnormal data points and then normalizes the remaining data. Then, the model performance is periodically evaluated during the model training process, and the network structure and learning weights are dynamically adjusted according to the evaluation results to adapt to changes in data distribution.

It is found that the defense mechanism can effectively reduce the interference of poisoning data on model training and improve the accuracy of the model on the poisoning dataset. By dynamically adjusting the network structure, the model can better adapt to the complexity of the data, while the adaptive learning weights help the model pay more attention to the normal data and reduce the learning of the poisoning data. In addition, the experimental results show that the mechanism exhibits better generalization ability in different datasets and attack scenarios.

Further research is needed in the future to investigate the effectiveness of this defense mechanism in larger datasets and more complex attack scenarios. In addition, exploring the combination of other defense strategies, such as data augmentation and model integration, may further enhance the robustness of the model. The significance of this study is to provide an effective defense method for machine learning models against data poisoning attacks, which helps to improve the security and reliability of the models in practical applications. With the increasing importance of data security and privacy protection, the results of this study have important practical applications for protecting machine learning models from malicious attacks.

References

1. Müller, N., Kowatsch, D., & Böttinger, K. (2020). Data Poisoning Attacks on Regression Learning and Corresponding Defenses. In 2020 IEEE 25th Pacific Rim International Symposium on Dependable Computing (PRDC) (pp. 80-89).

2. Ma, Y., Zhu, X., & Hsu, J. (2019). Data Poisoning against Differentially-Private Learners: Attacks and Defenses. In Proc. IJCAI (pp. 4732-4738).
3. Chen, J., Zhang, X., Zhang, R., Wang, C., & Liu, L. (2021). De-Pois: An Attack-Agnostic Defense against Data Poisoning Attacks. IEEE Transactions on Information Forensics and Security, 16, 34425.
4. Paracha, A., Arshad, J., Farah, M.B. et al. Deep behavioral analysis of machine learning algorithms against data poisoning. Int. J. Inf. Secur. 24, 29 (2025).
5. Y. Mao, D. Data, S. Diggavi and P. Tabuada, "Decentralized Optimization Resilient Against Local Data Poisoning Attacks," in IEEE Transactions on Automatic Control, vol. 70, no. 1, pp. 81-96, Jan. 2025
6. Z. Zheng, Z. Li, C. Huang, S. Long and X. Shen, "Defending Data Poisoning Attacks in DP-Based Crowdsensing: A Game-Theoretic Approach," in IEEE Transactions on Mobile Computing, vol. 24, no. 3, pp. 1859-1876, March 2025
7. Provably effective detection of effective data poisoning attacks. CoRR abs/2501.11795 (2025)
8. Alad, M., Catak, F., & Gul, E. (2019). Preventing Data Poisoning Attacks by Using Generative Models. 2019 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1-5).
9. Sponge Attack Against Multi-Exit Networks With Data Poisoning. IEEE Access 12: 33843-33851 (2024)
10. Ullah, S.S., Hussain, S., Ali, I. et al. Mitigating content poisoning attacks in named data networking: a survey of recent solutions, limitations, challenges and future research directions. Artif Intell Rev 58, 42 (2025).