A Machine Learning Framework For Data Poisoning Attacks

Priyanka Narsingoju[1] | Dr.D.Srinivas Reddy[2] |Dr.V.Bapuji[3]

[1]Department of MCA,Vaageswari College Of Engineering, Karimnagar.
[2]Associate Professor,Department of MCA,Vaageswari College Of Engineering, Karimnagar.
[3]Professor & HoD, Department of MCA,Vaageswari College Of Engineering, Karimnagar.

To Cite this Article
Priyanka Narsingoju, Dr.D.Srinivas Reddy ,Dr.V.Bapuji , "A Machine Learning Framework For Data Poisoning Attacks " *Journal of Science and Technology, Vol. 08, Issue 07,- July 2023, pp169-176*
Article Info
Received: 06-06-2023 Revised: 06-07-2023 Accepted: 17-07-2023 Published: 27-07-2023

ABSTRACT

Federated models are built by collecting model changes from participants. To maintain the secrecy of the training data, the aggregator has no visibility into how these updates are made by design.. This paper aims to explore the vulnerability of federated machine learning, focusing on attacking a federated multitasking learning framework. The framework enables resource-constrained node devices, such as mobile phones and IOT devices, to learn a shared model while keeping the training However, the communication protocol among attackers may take advantage of various nodes to conduct data poisoning assaults, which has been shown to pose a serious danger to the majority of machine learning models. The paper formulates the problem of computing optimal poisoning attacks on federated multitask learning as a bi-level program that is adaptive to arbitrary choice of target nodes and source attacking nodes.The authors propose a novel systems-aware optimization method, Attack confederated Learning(AT2FL), which is efficiency to derive the implicit gradients for poisoned data and further compute optimal attack strategies in the federated machine learning.

KEYWORDS: *Federated machine learning, Vulnerability,Arbitrary, Attack on federated machine learning(AT2FL), Gradients.*

INTRODUCTION

Machine learning has been widely applied in various applications, such as spam filtering and natural gas price prediction[1]. However, the reliability and security of these systems have been a concern, including adversaries. Researchers can rely on public crowd sourcing platforms or Private teams to collect training datasets, but both have the potential to be injected corrupted or poisoned data by attackers. It is crucial to research how well machine learning operates under poisoning it attempts in order to increase the resilience of real-world machine learning systems. Exploratory attacks and causal assaults are two categories of attack tactics. The n nodes in this federated learning system are shown by distinct colors. Corrupted or poisoned data is injected into certain nodes, whereas clean data is the sole data present in other nodes. The fundamental idea behind federated machine learning is to develop machine learning models based on data sets dispersed across numerous devices, while limiting data loss.

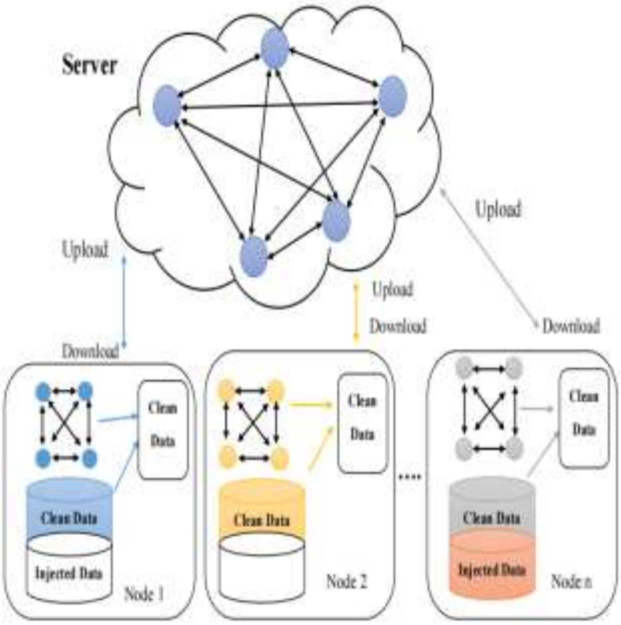


Fig.1 illustrates our data poisoning attack model for federated machine learning.

Although recent advancements have focused on overcoming statistical challenges (i.e., data collected across the network is in a nonrigid manner, with data on each node generated by a distinct distribution) or improving privacy preservation, attempts to make federated learning more reliable under poisoning attacks are still scarce. Consider multiple distinct e-commerce enterprises in the same region, and the goal is to develop a product purchase prediction model based on user and product information, such as the user's browsing and purchasing history. The attacker have access to a limited number of user accounts [2]. Furthermore, as a result of the existing communication mechanism between organizations, this protocol also allows the attacker to indirectly impact the inaccessible target nodes, which is also not addressed by existing poisoning techniques whose training data is gathered in a centralized place.

To analyze optimum poisoning attacks on federated machine learning in response to the preceding analysis. More precisely, as seen in figure 1,we concentrate on using the newly suggested federated multitask learning framework, a federated learning framework that records node interactions among numerous nodes in order to address statistical difficulties in a federated situation. The goal of our study is framed the optimal poisoning attack strategy on a federated multitask learning model as a universal bi-level improvisation problem that is adaptable to any combination of target nodes.

However, conventional optimization strategies for this bi-level issue are unsuitable for dealing with the system problems that arise in federated that offer a bi-level optimization framework for federated machine learning(e.g., high communication costs, stragglers). As a fundamental component of our study is to develop Attack on Federated Learning (AT2FL), a unique optimization approach for calculating poisoned data in the source attacking nodes. Furthermore, the generated gradient may be utilized to determine the best assault tactics[2].

Finally, the experimental test suggested optimum attack technique against random baselines on a variety of real-world datasets. The experiment results significantly validate our suggested model. Our suggested model is unique in three ways. To the best of our knowledge, this is an early attempt to investigate the vulnerability of federated machine learning from the standpoint of data poisoning.

To develop an efficient optimization approach, Attack on Federated Learning(AT2FL), for solving the optimal attack issue, which can handle system problems associated with federated machine learning. Using multiple real-world datasets, we illustrate the empirical performance of our optimum attack approach and our suggested AT2FL algorithm [10]. The experiment findings show that the communication protocol between several nodes allows attackers to assault federated machine learning.

I. Associated activity

The study is primarily concerned with data poisoning threats and federated machine learning, is to provide a quick overview of these two subjects. For data poisoning assaults, it has become an essential study subject in adversarial machine learning, where the target is machine learning algorithms. The previous attempt addresses poisoning attacks on support vector machines (SVM), where the selected attack employs a gradient ascent technique in which the gradient is derived depending on attributes of the SVM's best solution[9]. Furthermore, the poisoning attack is being researched on a variety of machine learning models, including auto regressive models, matrix factorization-bases collaborative filtering, and neural networks for graph data. In addition to single task learning models, maybe is the most relevant work to ours in the context of data poisoning assaults, as it is the first investigation on a considerably more difficult topic, namely the susceptibility of multitask learning [9].

However, the motives behind our work is markedly different. The data samples in are assembled, which differs from the scenario in federated machine learning, in which machine learning models are generated based on datasets disseminated across various nodes/devices while preventing data leakage.

The suggested algorithm is based on an optimization approach of multitask learning methods, which is unsuitable for dealing with the system issues in federated learning, such as large communication costs, and so on. Handling these issues in the context of data poisoning assaults is an important aspect. The basic goal of federated machine learning is to update classifiers quickly for current big datasets, and the training data it can handle has the following features.

1) Nonrigid: Each nodes/devices data may come from a different distribution.

2) Unbalanced: The amount of training samples differs by orders of magnitude for various nodes/devices. Federated learning can be classified according to the dispersion features of the data.

1. Horizontal (sample-based) federated learning, in which datasets share the same feature space but have distinct samples. A multitask type federated learning system is presented to allow many nodes to accomplish separate tasks while maintaining security and exchanging information.

2. Vertical federated learning, in which two datasets share the same sample ID space but differ in feature space.

Several privacy-preserving machine learning algorithms for vertically partitioned data have been proposed, including secure linear regression, gradient descent methods, and federated transfer learning, in which two datasets differ not only in samples but also in feature space. With the federated environment, classic transfer learning techniques may be used to generate solutions for the full sample and feature space. Presents a new model replacement methodology that leverages these flaws and demonstrates its usefulness on federated learning tasks as a first effort. Its goal, however, is to maintain high accuracy on the backdoor sub-task after assaulting. In contrast, By studying a poisoning assault against horizontal federated machine learning.

II. EXPERIMENTAL RESULT

The experimental assessment reported in the following sections shows the behavior of our suggested technique on an artificial two-dimensional datasets as well as its efficacy on the traditional MOIST

handwritten digit recognition datasets.3.1. Man-made data We begin with a two-dimensional data generation model in which each class has a Gaussian distribution and the mean and covariance matrices are = [1.5, 0], + = [1.5, 0], = + = 0.6I [1]. The points from the negative distribution are labeled 1 (shown in red in the
following figures), whereas the points from the positive distribution are labeled (shown in blue). The training and validation sets, D try and D val (25 and 500 points per class, respectively), are selected at random from this distribution.

The attacking class in the experiment shown below is the red one. In order to do this, a random point of the blue class is chosen and used as the beginning point for our procedure, with its label reversed. Then, until its termination condition is met, this assault is improved using our gradient ascent algorithm. Both the linear kernel's (upper two plots) and the RBF kernel's (lower two plots) attack trajectory is depicted in Figure 1 as a black line. Each plot's backdrop is an explicit computation of the error surface for every point inside the box
x [5, 5]2.

The rightmost plots in each pair show the classification error for the study region, whereas the leftmost plots in each pair display the hinge loss calculated on a validation set. For the linear kernel, the attack point range is constrained to the box x [4,4]2, which is shown as a dashed line. These figures demonstrate that for both kernels, our gradient ascent approach locates a respectable local maximum of the non-convex error surface[3]. Due to the unbounded nature of the error surface, the linear kernel ends at the corner of the bounded region. The hinge loss is also shown to have a nice local maximum for the RBF kernel, which also happens to be the largest classification error. Actual data using a well-known MOIST handwritten digit classification problem, we statistically assess the efficacy of the suggested assault technique.

To concentrate on two-class sub-problems of distinguishing between two different digits, much as Roberson camp; Rowers (2006).1We focus on the following two-class issues in particular: 7 vs 1, 9 versus 8, and 4 versus 0 are the results. A semantic meaning for an assault is provided by the visual depiction of writing digit data.

The MOIST data collection has each digit correctly normalized and rendered as a 28 by 28 pixel grayscale picture. In a raster1, each pixel is specifically organized.

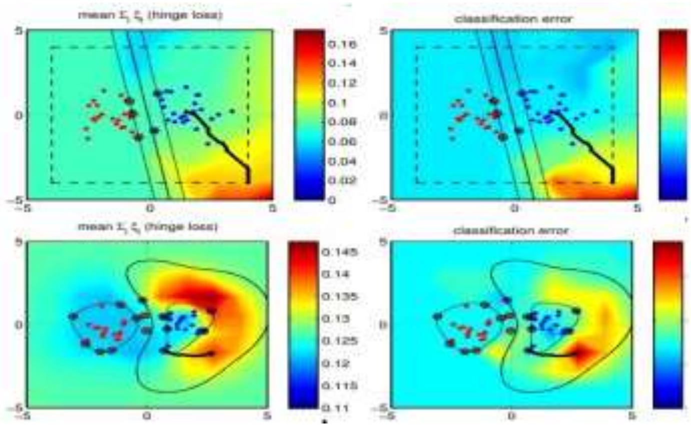


Figure 2. Displays the experiment's findings.

Behavior of the gradient-based attack strategy on the Gaussian data sets, for the linear (top row) and the RBF kernel (bottom row) with x = 0.5. The regularization parameter C was set to 1 in both cases. The solid black line represents the gradual shift of the attack point x (p) c toward a local maximum.

The hinge loss and the classification error are shown in colors, to appreciate that the hinge loss provides a good approximation of the classification error. The value of such functions for each point x

= [5, 5]2 is computed by learning an SVM on D tr {x, y = 1} and evaluating its performance on D val The SVM solution on the clean data D tr, and the training data itself, are reported for completeness, highlighting the support vectors (with black circles), the decision hyperplane and the margin bounds (with black lines).Scan and its worth are immediately regarded as features. D = 28 x 28 = 768 features make up the total amount of features. By dividing each feature's value by 255, able to normalize each pixel value between [0, 1]. Only the linear kernel is taken into account in this experiment, and C = 1 is used as the SVM's regularization parameter.

To perform the whole testing data provided by MOIST for DTS and randomly pick training and validation data of 100 and 500 samples, respectively. The testing data size is around 2000 samples per class (digit), albeit this number varies for each digit. Figure 2 displays the experiment's findings.

The example of the attacked class used as the starting point for our approach is displayed in the leftmost plots of each row. The ultimate

attack position is depicted in the middle plots. The graphic on the right shows how validation and testing mistakes rise as the attack goes on. The assault blurs the initial prototype towards the look of representatives of the attacking class, as seen by the attack point's visual appearance. The bottom segment of the 7 straightens out to resemble a 1, the lower segment of the 9 gets more rounded to resemble an 8, and round noise is added to the outside border of the 4 to make it resemble a 0. Comparing the original and final attack locations, we notice this impact. The rightmost charts clearly demonstrate how the mistake rate increased throughout the attack. Due to the reduced sample size, the validation error typically overestimates the classification error. However, a single assault data point in the sample runs described in this experiment led the classification error to increase from the early mistake rates of 2-5% to 15-20%. The mistakes in the first iteration of the rightmost plots shown in Figure 2 are caused by single random label flips, since our starting attack point is acquired by flipping the label of a point in the attacked class [1].

This underlines the SVM's susceptibility to poisoning assaults and shows that our approach can produce considerably larger mistake rates than random label flips. The latter point is further demonstrated in an experiment with several points and runs that is shown in Figure 3. The assault was lengthened in this experiment by using randomly selected training and validation sets of the same size (100 and 500 samples, respectively), adding extra points to the same class, and averaging outcomes over numerous runs.

With a rising percentage of attack points in the training set, it is evident that the attack effectiveness is steadily improving. The comparatively modest sizes of the training and validation data sets help to explain why the error variation is rather significant.

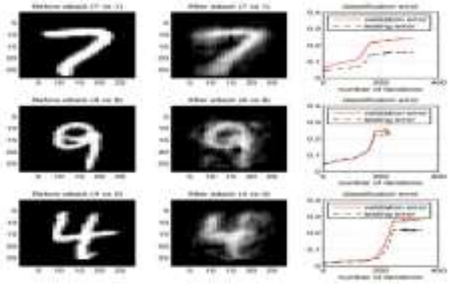
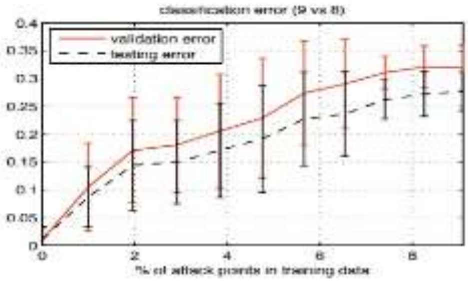


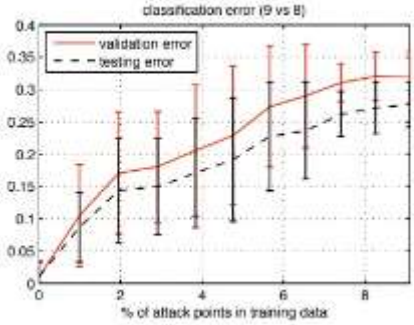
Figure 3: Shows how the proposed attack technique for the three two-class issues from the MOIST data set that were taken into consideration modified the original (mislabeled) assault point.

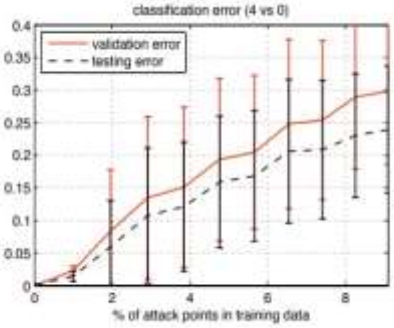
It is also noted that testing and validation faults have increased with successive revisions.

The latter point is further illustrated in a multiple point, multiple run experiment presented in Fig. 3. For this experiment, the attack was extended by using randomly selected training and validation sets of the same size (100 and 500 samples, respectively), adding extra points to the same class, and averaging outcomes over numerous runs. With a rising percentage of attack points in the training set, it is evident that the attack effectiveness is steadily improving. The comparatively modest sizes of the training and validation data sets help to explain why the error variation is rather significant. The poisoning assault described in this study serves as the starting point for an investigation of SVM's security against attacks on training data. Although it is debatable a simple algorithmic technique, our gradient ascent method has a surprisingly significant influence on the empirical classification accuracy of the SVM. The attack strategy described here also makes it possible to use differential operators to determine how modifications made to the input space affect the functions formed in the reproducing kernel Hilbert spaces. In contrast to other research on learning algorithm evasion (e.g., Bruckner camp; Schaeffer, 2009; Loft camp; Markov, 2010), this impact may make it easier to implement diverse evasion tactics in practice. There is still more to learn about these consequences. Future work needs to investigate a number of potential upgrades to the technique that is now being offered. The first would be to deal with the restriction of our optimization approach to tiny modifications in order to retain the structural restrictions of the SVM. By doing several small gradient steps, we solve. Investigating a method for computing the greatest step that can be taken with maintaining the structure of the ideal solution would be fascinating.

The simultaneous optimization of multi point assaults is an additional area for study, which we successfully tackled with sequential single-point attacks. The first concern is how to best disrupt a portion of the training data; in other words, one may generate simultaneous steps for each assault point to better optimize their combined effect rather than separately optimizing each attack point.







To pick the ideal subset of locations to utilize as the attack's launching point is the second issue. The latter is often a subset selection problem, however heuristics may provide better approximations.

Nevertheless, we show that the performance of the SVM is dramatically decreased by even subpar multi point attack techniques[3].

The assumption that the attacker has control over the labels of the injected points is a significant practical constraint of the proposed strategy. These presumptions might not be true if labels are solely given by reliable sources, like people. For instance, a spam filter bases its decisions on how users have classified communications. As a result, even though an attacker can send any message, he cannot ensure that it will include the labels required for his assault.

This imposes the additional restriction that, in order to trick the labeling oracle, the attack data must fulfil specific side constraints.

Understanding these possible side limitations and using them into assaults will require more research. Incorporating the real-world inverse feature-mapping issue, or the challenge of locating real-world attack data that may produce the required outcome in the learner's input space, would be the ultimate addition. There is a direct mapping between the input characteristics used for learning and the real-world picture data for data like handwritten numbers.

The mapping is more complicated and may include several non-smooth operations and normalization in many other issues (such as spam filtering). These inverse mapping puzzles for learning assaults have not yet been solved.[1]

III. CONCLUSION

To discover and assessed a fresh flaw in federated learning. Federated learning provides hundreds or even millions of participants, some of whom may unavoidably be evil, with direct control over the weights of the jointly learned model through model averaging. This makes it possible for a malicious participant to add a backdoor sub task to the shared model.

Since secure aggregation is used to keep participant non-i.i.d. local training data private and federated learning is meant to benefit from it, anomaly detection cannot be implemented and would not have been effective anyhow. We created a brand-new model-replacement approach that takes use of these flaws and proved its effectiveness on common federated-learning tasks like word prediction and picture categorization. Even when previously suggested data poisoning techniques fail or require many malevolent players, model substitution successfully injects backdoor.

The huge capacity of current deep learning models is another aspect that helps backdoor assaults succeed. Traditional measurements of model quality do not account for the model's additional learning; they only assess how effectively it has learned its primary task. The model's accuracy won't be significantly impacted if covert backdoor are added using this extra capacity. Federated learning goes beyond being a distributed implementation of conventional machine learning.

Due to the dispersed nature of the system, it must be resilient to players who act inappropriately on a whim. Unfortunately,

conventional methods for Byzantine-tolerant distributed learning do not work when which is precisely the situation that spurs the development of federated learning. The training data of the participants are not uniquely identified,which is precisely the situation that spurs the development of federated learning[1].

IV. REFERENCES

[1]. Alfeld, S., Zhu, X., & Barford, P. (2016). Data Poisoning Attacks against Autoregressive Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1). https://doi.org/10.1609/aaai.v30i1.10237

[2]. Vitaly Shmatikov, Yiqing Hua, Deborah Estrin, Eugene Bagdasaryan, and Andreas Veit. federated learning backdoor techniques. 2018. arXiv preprint arXiv:1807.00459.
https://doi.org/10.48550/arXiv.1807.00459

[3]. Barreno, M., Nelson, B., Joseph, A.D. *et al.* The security of machine learning. *Mach Learn* 81, 121–148(2010). https://doi.org/10.1007/s10994-010-5188-5

[4]. Zhang S ,Wang C and Zomaya C. Robustness Analysis and Enhancement of Deep Reinforcement Learning-Based Schedulers. *IEEE Transactions on Parallel and Distributed Systems*. 10.1109/TPDS.2022.3218649. 34:1. (346-357
https://ieeexplore.ieee.org/document/9937194/

[5]Pavel Laskov, Blaine Nelson, and Batista Biggio. assaults with poison on support vector machines.
2012's arXiv preprint 1206.6389
https://doi.org/10.48550/arXiv.1206.6389

[6]Andrew Zisserman, Karen Simonyan, Ken Chatfield, and Andrea Vedaldi.
The devil is once again in the details: a thorough examination of convolution networks.
2014 arXiv preprint 1405.3531.
https://doi.org/10.48550/arXiv.1405.353

[7]Massimo Fornasier, Ignace Loris, and Ingrid Daubechies. Accelerated projected gradient approach for sparsity-constrained linear inverse problems. 2008;14(5-6):764–792;Journal of Fourier Analysis and Applications.
https://doi.org/10.48550/arXiv.0706.4297

[8] Mariana Raykova, Phillip Schoppmann, and Adrià Gascón. On datasets that have been vertically partitioned, secure linear regression.
https://doi.org/10.48550/arXiv.0706.4297

[9] S. Han, W. K. Ng, L. Wan and V. C. S. Lee, "Privacy-Preserving Gradient-Descent Methods," in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 6, pp. 884-899, June 2010, doi: 10.1109/TKDE.2009.153.

[10]Ling Huang, JD Tygar, Anthony D. Joseph, Blaine Nelson, and Benjamin I. P. Rubinstein. Machine learning that is hostile. Pages 43–58 in the book Proceedings of the Fourth ACM Workshop on Security and Artificial Intelligence.
ACM, 2011.
https://ieeexplore.ieee.org/document/9887796/