SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7106361

INVESTİGATİON ON DETECTİON OF DATA POİSONİNG ATTACKS: MOST POSSİBLE DEFENCES AND COUNTER MEASURES

Kireet Muppavaram
Assistant Professor, Dept of CSE, School of Technology, GITAM DEEMED TO BE University Hyderabad, Email : kmuppava@gitam.edu

Aparna Shivampeta
Assistant Professor, Dept of CSE, School of Technology, GITAM DEEMED TO BE University Hyderabad. Email : ashivampeta@gitam.edu

Hyma Biruduraju
Assistant Professor, Dept of CSE, Gurunanak Institutions Technical campus, Hyderabad
Email : hymaomkaram@gmail.com

Vishwesh Nagamalla
Assistant Professor, Dept of CSE, Sreenidhi Institute of Science and Technology, Hyderabad
Email : vishwesh2010@gmail.com

Ishmatha begum
Assistant Professor, Dept of CSE, Gurunanak Institutions Technical campus, Hyderabad
Email : ishmathabegum.gnit@gniindia.org

Abstract. Machine learning has become one of the most prominent applications in various fields for the development of high end systems. This trend of machine learning applications usage made the attackers to choose the machine learning applications models and induce different type of attacks like data poisoning attacks, adversarial attacks, Obfuscation Attacks, Side channel attacks, Model Inversion attacks, MITM attacks. It is very essential to provide security to the machine learning models by protecting the integrity, confidentiality and availability of the training data, testing data of machine learning models. Through our study we found that the data poisoning attacks are the majority of the attacks attempted by the attackers on machine learning systems. In this paper we carefully analysed data poisoning attacks from the existing models and by our investigation we proposed the most possible defences and countermeasures to Data Poisoning attacks.

Keywords: Machine learning, Data poisoning attacks, integrity, confidentiality, availability, brute force attacks.

1 Introduction

Today machine learning has become a most prominent application in the development of AI systems. The trend of Artificial Intelligence all over the world gave rise to the development of

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7106361

different algorithms to train the machines. This vast usage of machine learning applications gave a platform for the attackers to introduce the attacks using the machine learning applications. To ensure security from the attacks on machine learning applications, "Security in Machine Learning" has become a major area to work on for the researchers.

In this current machine learning era, machine learning security has grabbed the attention all over the world. The dependency of automated systems using machine learning systems is increasing; as a result, security in machine learning becomes the need of the day. The security on machine learning applications or systems can be ensured by considering the basic security goals achieved by the system. The security goals [2] i.e. CIA (confidentiality, integrity and availability) ensures the security level of the machine learning application.

In this paper we analysed the various attacks which violated security goals on machine learning systems. In our analysis we tried to identify the entry point of attacks and we found few major entry points by which these attacks can be initiated on machine learning systems. The identified major entry points are (i) attacks using training data (ii) attacks by duplicating models (iii) MITM attacks. The attacks using training data are considered as integrity breach, attacks by duplicating models considered as confidentiality and availability breach and MITM attacks are considered as integrity breach.

We analysed the following attacks from the previous researches and identified the major attacks. The attacks are (i) Data poisoning attacks [3] [4] [5], (ii) adversarial attacks [6][7][8], Obfuscation Attacks [9] [10], Side channel attacks [11], Model Inversion attacks [12], MITM attacks [13]. In this analysis we found that majority of the attacks on machine learning systems are done using Data Poisoning attacks using different standard reports [28]. By considering this reports we carefully investigated the data poisoning attacks.

This paper is organized in the following way, Section 2 provides the related work in security in machine learning, Section 3 provides Poisoning attacks and attacker abilities Section 4 provides the most possible defences and countermeasures to data poisoning attacks. Finally, section 5 provides the conclusion.

2 Related Work

Chenglin Miao et al [14] worked on data poisoning attacks by attackers in crowd sensing systems and proposed an enhanced mechanism to reduce the data poisoning attacks. They used two attacks in their proposed work i.e. availability attack, target attack and built a practical approach for crowd sensing system evaluation.

Kui Ren et al [15] worked on adversarial attacks and investigated threat models which distinguishes black-box, white-box and gray-box attacks. In black box threat model challenger depends only on query access, in gray box threat model the challenger depends only structure of target model. In whitebox threat model the challenger gains full knowledge of target model. This paper has limitations as the threat method requires large number of queries.

Sebastian Banescu et al [16] worked on Obfuscation Attacks and proposed an approach for extracting program features that are prevalent in predicting the automated attacks for protecting the software. They built a test regression model based on symbolic execution in order to predict the obfuscation attacks. In their approach the main limitation is "lack of space" so they

DOI: 10.5281/zenodo.7106361

represented with limited parameters to obtain better results.

Maria Mushtaq et al [17] worked on detection of side channel attacks by examining the usage of machine learning techniques on Intel x86 architecture to detect Cache based side channel attacks. Finally, they produced the minimum selection metrics used for machine learning techniques in order to carry-out run-time Cache based side channel attacks detection in real time scenario.

Seira Hidano et al [18] worked on Model Inversion attacks and proposed a general model inversion framework. Their work concentrated towards extracting the supplementary information which is available to the challenger. This paper also shows that sensitive attributes can be gathered by mining non sensitive attributes which modifies the machine learning model into targeted model using the techniques of data poisoning. This models limitation is it requires previous distribution p as supplementary information.

Cheng-Yu Cheng et al [19] worked on MITM attacks and proposed a model which uses network packet analysis, techniques in machine learning to calculate the difference in packet Round-trip-time (RTT) between user and receiver. The limitation in this model is if the attacker uses wired connection and client uses the wireless connection then it would be difficult to calculate the RTT.

M. Aladag, et al [29] This work shows how an attacker can access the data by using manipulation i.e. the attacker manipulates abnormal behaviour as normal behaviour The limitation of this model is that this model is much dependent on auto-encoder model.

N. Baracaldo et al [30] This work more concentrates towards detecting the IoT devices. The proposed method is a novel method for detecting and also filtering the poisonous data in order to train the supervised learning models which is suitable for IoT environments.

3 Salient Findings from analysis

In this study of analysing the attacks done on machine learning algorithms we found that the data poisoning attacks are the major attacks in machine learning systems by which the attacker attacks the training models which affect the entire machine learning process. There are different research efforts made by the different researchers [19] [20] [21] [22] in reducing this attacks but still these data poisoning attacks stands out as the major research work till date.

We carefully analysed the data poisoning attacks and found the attacker capabilities to attack the machine learning systems. Data poisoning attacks can be done by the attackers in two ways (i) data poisoning before training the model (ii) data poisoning post building of the model. An attacker uses different type of tricks or attempts to induce the poisoning into the training data of machine learning system. The following are the different type of attempts or abilities of the attacker to attack machine learning systems they are Data manipulation, Logic corruption, Transfer learning, Data injection.

Data manipulation: In this type of attack, attacker will manipulate the training data by modifying, removing or adding the data to the trained datasets. Considering the scenarios of

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7106361

new labels in the training models attackers try to gain the information and change the labels using random label flipping or heuristics.

Logic corruption: In this type of attack, attacker modifies the algorithm which is used to train the machine. This is the most dangerous attack as it completely changes the mode of the machine.

Transfer learning: In this type of attack, attacker gains the information related to the model which is reused as transfer learning model for different machines. These types of attacks are also called as MITM attacks.

Data injection: In this type of attack attacker injects the data into training datasets and changes the mode of the training model. This kind of attack is similar to data manipulation where data manipulation attacks are more concentrated towards modifying the training labels, here in data injection attacks data inside the trained datasets are modified added or removed.

4 Defences and Countermeasures to Data poisoning attacks

In this study we analysed the most common way used by the attackers to attacks machine learning systems using data poisoning attacks.

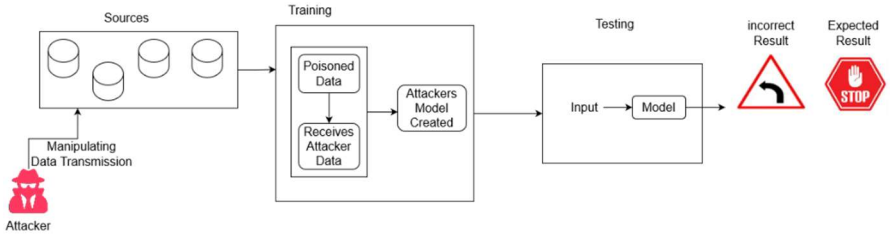

Fig. 1. Model for attacking machine learning system

It can be observed from the figure that attackers in the first phase try to gather information using various sniffing tools and retrieve the information which can be useful for attacking the system. The information can be related to trained datasets. Finally, the action plan is performed by the attacker to attack in either of the two ways (i) data poisoning before training the model (ii) data poisoning post building of the model. Data poisoning attacks are the attacks which are dependent on the knowledge gained by the attackers in the initial phase.

Attackers abilities to gain Knowledge: Here, considering the views of the (i) attacker can attack as (A1) by modifying the retrieved trained datasets, (ii) attackers can attack as (A2) by modifying the algorithm (iii) attackers can attack as (A3) by brute force attack.

SEYBOLD Report
ISSN: 1533 - 9211
DOI: 10.5281/zenodo.7106361

Attacker attacking capabilities:

- ✓ **Black-box attacks:** The attackers does not gain any information related to trained datasets but performs the attacks using the brute force attacks.
- ✓ **Graybox attacks:** The attackers perform this type of attacks using the gained information or the brute force attacks.
- ✓ **Whitebox attacks:** The attacks gain the information related to trained datasets, algorithms and applies few injection techniques.

Attack Specificity: This refers to the exact data points targeted by the attacker.

Defences:

By analysis, it came to know that attacker can attack by A1, A2 and A3 ways.

- **Defence for A1:** Considering these 3 scenarios from the above the input features which forms a collection for training datasets needs to be verified. The possible way of verifying the input features collection can be done by assigning the weights to the input. By considering weighted values cost of each input feature can be obtained which can be used to detect the actual data collection required for the training datasets.
- **Defence for A2:** In order to check whether the algorithm is modified or not. A checksum is generated to the specific algorithm and applied. In verification process whether the algorithm has been modified or not checksum matching is used.
- **Defence for A3:** Brute force attacks can be avoided by testing the network flow [27]. The collected data is considered as labeled by professional network experts and then each flow is identified whether it leads to a brute force attack.

5. Conclusion

The enormous usage of machine learning models in the present society has made this area popular in various scientific and research purposes. This Popular usage of machine learning systems has become a platform for the attackers to attack by introducing malicious activities using machine learning systems. In this paper we analysed the different attacks done on machine learning models and we found that the data poisoning attacks are the major attacks in machine learning systems by which the attacker attacks the training models which affect the entire machine learning process. By our careful investigation on data poisoning attacks we presented the most possible defences and countermeasures to data poisoning attacks.

References

1. M. Xue, C. Yuan, H. Wu, Y. Zhang and W. Liu, "Machine Learning Security: Threats, Countermeasures, and Evaluations," in IEEE Access, vol. 8, pp. 74720-74742, 2020, doi: 10.1109/ACCESS.2020.2987435.
2. Popescul, Daniela. The Confidentiality – Integrity – Accessibility Triad into the Knowledge Security. A Reassessment from the Point of View of the Knowledge Contribution to

DOI: 10.5281/zenodo.7106361

Innovation. Proceedings of The 16th International Business Information Management Association Conference (Innovation and Knowledge Management, A Global Competitive Advantage), June 29-30, 2011, Kuala Lumpur, Malaysia, Editor Khalid S. Soliman, ISBN: 978-0-9821489-5-2, pp. 1338-1345

3. Battista Biggio, Samuel Rota Bulo, Ignazio Pillai, Michele ` Mura, Eyasu Zemene Mequanint, Marcello Pelillo, and Fabio Roli. "Poisoning Complete-Linkage Hierarchical Clustering". In: ed. by Ana Fred, Terry M. Caelli, Robert P. W. Duin, Aurelio C. Campilho, and Dick de Ridder. ´ Vol. 3138. Springer Berlin Heidelberg, Aug. 2004. ISBN: 978-3-540-22570-6. DOI: 10 . 1007 / b98738. arXiv: 9780201398298
4. Battista Biggio, Ignazio Pillai, Samuel Rota Bulo, Davide ` Ariu, Marcello Pelillo, and Fabio Roli. "Is data clustering in adversarial settings secure?" In: Proceedings of the 2013 ACM workshop on Artificial intelligence and security - AISec '13 (2013), pp. 87–98. ISSN: 15437221. DOI: 10.1145/2517312.2517321
5. Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K. Jha. "Systematic poisoning attacks on and defenses for machine learning in healthcare". In: IEEE Journal of Biomedical and Health Informatics 19.6 (2015), pp. 1893–1905. ISSN: 21682194. DOI: 10 . 1109/JBHI.2014.2344095.
6. Abigail Graese, Andras Rozsa, and Terrance E. Boult. "Assessing threat of adversarial examples on deep neural networks". In: Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016 (2017), pp. 69–74. DOI: 10.1109/ICMLA. 2016.44. arXiv: 1610.04256.
7. Jamie Hayes and George Danezis. "Machine Learning as an Adversarial Service: Learning Black-Box Adversarial Examples". In: (2017). arXiv: 1708.05207
8. Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. "Adversarial Attacks on Neural Network Policies". In: (Feb. 2017). arXiv: 1702.02284
9. Battista Biggio, Ignazio Pillai, Samuel Rota Bulo, Davide ` Ariu, Marcello Pelillo, and Fabio Roli. "Is data clustering in adversarial settings secure?" In: Proceedings of the 2013 ACM workshop on Artificial intelligence and security - AISec '13 (2013), pp. 87–98. ISSN: 15437221. DOI: 10.1145/2517312.2517321.
10. Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger, Igino Corona, Giorgio Giacinto, and Fabio Roli. "Poisoning behavioral malware clustering". In: Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop - AISec '14. New York, USA: ACM Press, Nov. 2014, pp. 27–36. ISBN: 9781450331531. DOI: 10.1145/2666652.2666666.
11. Lingxiao Wei, Yannan Liu, Bo Luo, Yu Li, and Qiang Xu. "I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators". In: (2018). arXiv: 1803.05847
12. Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures". In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15. New York, USA: ACM Press, 2015, pp. 1322–1333. ISBN: 9781450338325. DOI:

DOI: 10.5281/zenodo.7106361
10.1145/2810103.2813677

13. D. Wang, C. Li, S. Wen, and Y. Xiang "Man-in-the-Middle Attacks against Machine Learning Classifiers via Malicious Generative Models" School of Software and Electrical Engineering, Swinburne University of Technology, Hawthorn, VIC 3122, Australia IEEE TDSC, OCTOBER 2019
14. C. Miao, Q. Li, H. Xiao, W. Jiang, M. Huai, and L. Su, “Towards data poisoning attacks in crowd sensing systems,” in Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput. (Mobihoc), 2018.
15. Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu, Adversarial Attacks and Defenses in Deep Learning, Engineering, Volume 6, Issue 3,2020, Pages 346-360, ISSN 2095-8099, https://doi.org/10.1016/j.eng.2019.12.012. http://www.sciencedirect.com/science/article/pii/S209580991930503X)
16. Sebastian Banescu, Technische Universität München; Christian Collberg Predicting the Resilience of Obfuscated Code Against Symbolic Execution Attacks via Machine Learning" 26th USENIX Security Symposium August 16–18, 2017 Vancouver, BC, Canada ISBN 978-1-931971-40-9 .
17. M. Mushtaq et al., "Machine Learning For Security: The Case of Side-Channel Attack Detection at Run-time," 2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS), Bordeaux, 2018, pp. 485-488, doi: 10.1109/ICECS.2018.8617994.
18. HIDANO, Seira & Murakami, Takao & KATSUMATA, Shuichi & Kiyomoto, Shinsaku & Hanaoka, Goichiro. (2018). Model Inversion Attacks for Online Prediction Systems: Without Knowledge of Non-Sensitive Attributes. IEICE Transactions on Information and Systems. E101.D. 2665-2676. 10.1587/transinf.2017ICP0013.
19. Battista Biggio, Konrad Rieck, Davide Ariu, Christian Wressnegger "Poisoning Behavioral Malware Clustering " 2014 ACM CCS Workshop on Artificial Intelligent and Security, AISec '14, pages 27-36, New York, NY, USA, 2014. ACM
20. Huang Xiao and Battista Biggio and Gavin Brown and Giorgio Fumera and Claudia Eckert and Fabio Roli "Is Feature Selection Secure against Training Data Poisoning?" Proc. of the 32nd ICML, Lille, France, 2015. JMLR: W&CP vol. 37
21. Andrew Newell, Rahul Potharaju, Luojie Xiang, and Cristina Nita-Rotaru " On the Practicality of Integrity Attacks on Document-LevelSentiment Analysis" : AISec '14: Proceedings of the 2014 Workshop onArtificial Intelligent and Security Workshop November 2014 Pages 83–93 https://doi.org/10.1145/2666652.2666661
22. Benjamin I. P. Rubinstein, Blaine Nelson, Ling Huang " ANTIDOTE: understanding and defending against poisoning of anomaly detectors" MC '09: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement November 2009 Pages 1–14 https://doi.org/10.1145/1644893.1644895
23. M.Kireet, Pavithra rachala, Dr.Meda Sreenivasa Rao, Rukmini Sreerangam ”Investigation of Contemporary Attacks In Android Apps” published in International Journal of scientific and Technology Research (IJSTR),Volume 8, issue 12,December 2019, ISSN 2277-8616

DOI: 10.5281/zenodo.7106361

24. M. Kireet, Dr. Meda Sreenivasa Rao. “Investigation of Collusion Attack Detection in Android Smartphones.” International Journal of Computer Science and Information Security, (IJCSIS) Vol. 14, No. 6, June 2016.
25. Kireet .M,Dr.Meda Sreenivasa Rao “A Survey on Malware attacks on smartphones (IJCSIT) “ ISSN : 09759646 Volume 6 issue 3 2015.
26. Muppavaram K., Sreenivasa Rao M., Rekanar K., Sarath Babu R. (2018) How Safe Is Your Mobile App? Mobile App Attacks and Defense. In: Bhateja V., Tavares J., Rani B., Prasad V., Raju K. (eds) Proceedings of the Second International Conference on Computational Intelligence and Informatics. Advances in Intelligent Systems and Computing, vol-712. Springer,Singapore Print ISBN 978-981-10-8227-6 DOI: https://doi.org/10.1007/978-981-10-8228-3_19 2018.
27. M. M. Najafabadi, T. M. Khoshgoftaar, C. Kemp, N. Seliya and R. Zuech, "Machine Learning for Detecting Brute Force Attacks at the Network Level," 2014 IEEE International Conference on Bioinformatics and Bioengineering, Boca Raton, FL, USA, 2014, pp. 379-385, doi: 10.1109/BIBE.2014.73.
28. Standard Reports on Data Posoning attacks Website: https://www.kaspersky.com/enterprise-security/wiki-section/products/machine-learning-in-cybersecurity
29. M. Aladag, F. O. Catak and E. Gul, "Preventing Data Poisoning Attacks By Using Generative Models," 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 2019, pp. 1-5, doi: 10.1109/UBMYK48245.2019.8965459.
30. N. Baracaldo, B. Chen, H. Ludwig, A. Safavi and R. Zhang, "Detecting Poisoning Attacks on Machine Learning in IoT Environments," 2018 IEEE International Congress on Internet of Things (ICIOT), San Francisco, CA, USA, 2018, pp. 57-64, doi: 10.1109/ICIOT.2018.00015.