Countermeasures in Latent Attribute Inference against Data Poison Attacks

Albert Yen
Department of Cybersecurity
New York University
New York, NY
Acy238@nyu.edu

***Abstract*—An adversary can manipulate training data to cause a machine learning model to learn abnormal behavior as normal behavior. We will examine the robustness of Logistic Regression models against a specific data poison attack, then propose a defense to mitigate the effects of poisoned data points through latent attribute inference. After testing our hypothesis, we find inference from neighborhood feature vectors as an adequate defense, which can be expanded upon in future research.**

***Keywords*—homophily, logistic regression, data poison attack**

I. Introduction

The eventual convergence of cybersecurity and data science in social media is inevitable – this confluence has also sparked a new awakening in public interest with data privacy. From AOL to the more recent Facebook and Cambridge Analytica, the practice of data mining on social media has garnered attention towards data governance and privacy policies. The primary focus of this paper is not on confidentiality nor availability of mined data, but rather on integrity, and the robustness of Machine Learning (ML) models that rely on social media training data, which are often sold as boutique datasets.

Past research demonstrated how data poisoning attacks that introduce varying degrees of noise to the training data – even 5% noise – can create a feature imbalance that decreases classification accuracy [1] [2]. The broader implications of such attacks will eventually trickle down into the prosaic technologies that pervade our daily lives; from political campaign mailers to Support Vector Machine (SVM) classifiers for cancer diagnosis, any technology that relies on classification accuracy will suffer in performance. Our research examines the effects of various data poisoning levels on Logistic Regression (LR) models that predict whether a user identify as Democrat, and subsequently propose a defense against such attacks. We hope our findings can be extended to future research in other industries.

While there exists detection and filtration tools for mining data, the practice and ethics of data mining are beyond the scope of our paper since we focus on mined data integrity. Our paper is organized as follows: Section II showcases related works. Section III briefly examines our adversary's motivations and attack options. Section IV explores a case study, proposes our hypothesis, and measures our success rate. Lastly, Section V will summarize our results and conclude with a potential direction for future research.

II. Related Works

Aladag et al. proposed a neural network with a binary cross entropy loss function that learns the natural structure of true data and fortifies a classification model's robustness against optimized data poisoning attacks [3]. While this proposal – an auto-encoder model with varying compression ratios and activation functions – demonstrates a level of resiliency from SVM against poisoning attacks, we propose a defense specifically for LR models.

Jagielski et al. proposed a new algorithm, TRIM, to strength the robustness of Linear Regression models against optimized data poisoning attacks [4]. TRIM iterates on a subset of datapoints to estimate the lowest residuals that will separate legitimate data points from attack points. This approach demonstrated promising results when compared to existing defense proposals and was evaluated in the following industries: healthcare, loan assessments, and real estate. To contrast, our research is solely focused on LR models within social media.

Alufaisan et al. showcased existing vulnerabilities in traditional data mining techniques, and then evaluated how well can poisoned data points deceive a model that predicts user traits from "likes" on Facebook [1]. From Feature-Altering attacks to Fake-Users Addition attacks, Alufaisan et al. proposed Flip on Negative Impact (FONI) defense against these various data poisoning techniques; and proposed Strong Ties Detection defense against Evasion attacks. While these countermeasures reduced the severity of said attacks for classification models – the results varied with changes in percentage of total data points poisoned. Our work builds on the research from Alufaisan et al. and looks to defend against data poisoning attacks.

III. Motivation and Attack Options

We assume our adversary is intrinsically motivated to maximize damage while minimizing the required computing costs for attacks, and our adversary has numerous options for data poisoning attacks. Due to page limitations, we will examine Class-Altering poison attacks – flipping of users from the opposite class to target class – on our LR model's training data. Next, we will calculate our benchmark LR model's accuracy and measure our adversary's attack efficiency before, and after, implementing our proposed countermeasure.

IV. Case Study and Proposal

We examine hashtags with political sentiment used in tweets as our feature vectors in training our LR model to predict whether a profile identifies as Democrat. We follow comparative preprocessing steps from Alufaisan et al. to prevent a feature imbalance, but we choose data from Twitter over that of Facebook since we prefer a network graph with direction. Users' neighborhood feature vectors will be formed by their friends and not followers, since users have more control on who they follow than who follows them and what they see than how they are seen.

As a countermeasure, we propose to extend latent attribute inference from precise neighborhood feature vectors upon negative data point impacts. Our hypothesis relies on the assumptions of homophily and how it is relatively easier for our adversary to control a profile's features than its entire network of friends' features. Zamal et al. have showcased neighborhood feature vectors alone are sufficient to match or surpass the inference accuracy than that of user-only feature vectors [5]. Their research emphasized the need for various neighborhood modifications in apposite feature contexts, i.e., closest friends for gender or all friends for politics.

In testing our hypothesis, we use a subset of data from SNAP that originated from Twitter [6]. When we consider the storage required for the computation of 400 profiles and their associated network connections, we require over 100GB of memory. For explicability and simplicity, we limit our data to 60 profiles. To prevent overfitting, we reduce our feature variables from 500 to 200 germane hashtags. We remove any profile that has less than two friends or tweets. We ensure an even distribution of profiles that identify as Democrat, 30, and non-Democrat, 30, for a total of 60 profiles.

With 200 political hashtags as our feature selection, the input space is modeled as an n-dimensional discrete value space $K = (X_1, ..., X_n)$, where χ_i represents the i–th feature. Each instance $\chi_i = (\chi_{i1}, ..., \chi_{in})$, where $\chi_{ij} \in X_j$, is mapped to a target class, Y. We assume binary domain for our target class $Y \in \{C_t, C_o\}$, where C_t represents our target class and C_o represents our opposite class. We choose stratified k-fold cross validation with 6 folds to validate our model – the figure below details our accuracy without any data poison through Area Under the ROC curve (AUC).

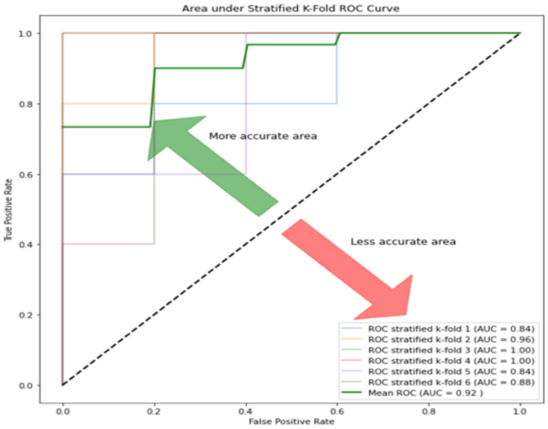

Figure 1. LR Model's Benchmark AUC

If our classification accuracy improves without χ_i in our training data, then we know this instance has a negative impact on accuracy. For said instances, we will extend latent attribute inference from precise neighborhood feature vectors to Y_i. The figures below calculate our LR model's accuracy with, and without, our proposal at various poison percentages.

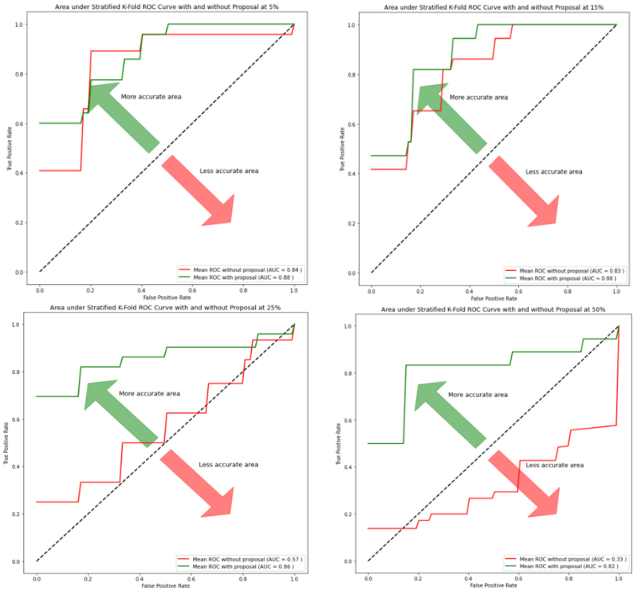

Figure 2. Proposal's success at various poison percentages

When compared to the original LR model's 0.92 AUC, poison attacks of various percentages were successful. Although the most successful poison attack was when 50% of our total training data was poisoned – AUC decreased from 0.92 to 0.33 without our proposed defense but improved to 0.82 with – our proposal demonstrates resiliency and improved AUC across all levels of poison. We resolved poisoned instances with our proposed defense on utilizing neighborhood data to decrease our adversary's attack efficiency.

Although our proposal did not completely eliminate data poison and meet our benchmark AUC, our proposal provided reliable and consistent defense that significantly reduced our adversary's attack efficiency across all percentages of data poison on our total training data. Our findings validate the results of Zamal et al. and suggest neighborhood data can improve accuracy. Our preliminary metrics proved neighborhood data can be extended as a defense against poison attacks on LR models.

V. Conclusion

Our LR model's training data included users' neighborhood data, i.e., network connections, which fortified robustness and defended against data poison attacks of all levels on our training data.

These findings suggest an abundance of inherent feature attributes within social media users' network connections, which can improve the classification accuracy than that of user-only feature vectors against data poison attacks. A potential direction for future research would be to measure various data compression approaches for network connections' memory storage with the least inadvertent effects on LR models, or other ML models.

Bibliography

[1] Y. Alufaisan, Y. Zhou, M. Kantarcioglu and B. Thuraisingham, "Hacking social network data mining," in *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Beijing, China, 2017.

[2] B. Thuraisingham, M. Kantarcioglu and L. Khan, "Integrating Cyber Security and Data Science for Social Media: A Position Paper," in *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, Vancouver, BC, Canada, 2018.

[3] M. Aladag, F. O. Catak and E. Gul, "Preventing Data Poisoning Attacks By Using Generative Models," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, Ankara, 2019.

[4] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru and B. Li, "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning," in *2018 IEEE Symposium on Security and Privacy (SP)*, San Francisco, 2018.

[5] F. A. Zamal, W. Liu and D. Ruths, "Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors," in *International Conference on Weblogs and Social Media (ICWSM)*, Montreal, Quebec Canada, 2012.

[6] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford Large Network Dataset Collection," Stanford SNAP, 2014.