International Journal of Cyber Warfare and Terrorism
Volume 10 • Issue 2 • April-June 2020

Insider Threat Detection Using Supervised Machine Learning Algorithms on an Extremely Imbalanced Dataset

Naghmeh Moradpoor Sheykhkanloo, Edinburgh Napier University, Edinburgh, UK
Adam Hall, Edinburgh Napier University, Edinburgh, UK

ABSTRACT

An insider threat can take on many forms and fall under different categories. This includes malicious insider, careless/unaware/uneducated/naïve employee, and the third-party contractor. Machine learning techniques have been studied in published literature as a promising solution for such threats. However, they can be biased and/or inaccurate when the associated dataset is hugely imbalanced. Therefore, this article addresses the insider threat detection on an extremely imbalanced dataset which includes employing a popular balancing technique known as spread subsample. The results show that although balancing the dataset using this technique did not improve performance metrics, it did improve the time taken to build the model and the time taken to test the model. Additionally, the authors realised that running the chosen classifiers with parameters other than the default ones has an impact on both balanced and imbalanced scenarios, but the impact is significantly stronger when using the imbalanced dataset.

KEYWORDS

Data Pre-Processing, Imbalanced Dataset, Insider Threat, Spread Subsample, Supervised Machine Learning

1. INTRODUCTION

Insider attacks present a considerable issue in the cyber-threat landscape, with 40% of organisations labelling the vector as the most damaging attack faced (Cole, 2017) and (Moradpoor, 2017). In 2016, the containment and remediation of reported insider threats cost affected organisations 4 million dollars on average (Ponemon Institute, 2016). In addition, insider threats are extremely common among cyber-incidents; in 2015, 55% of cyber-attacks were insider threat cases (Bradley, 2015). Despite the high cost and frequent occurrence of insider threat attacks, detection and mitigation remain a problem. In 2018, 90% of companies are regarded vulnerable (Insiders, 2018). A further 38% of companies acknowledge that their insider threat detection and prevention capabilities are not adequate (Cole, 2017). This disparity demonstrates a significant gap between the current advancements in insider threat detection, and the requirements of businesses. Given the availability of computational resources, it is feasible to use Machine Learning (ML) techniques to solve problems of larger complexity than has previously been possible. A strong precedent of this can be observed in recent history with the growth of the field of Big Data. This is also exemplified by the historic achievement of Google Deepmind (Hassabis, 2017), creating a machine learning algorithm which masters the immensely complex board game Go (Silver, 2016). Most organisations have the resources to keep logs of employee interactions with technology. By harnessing the data produced through logging, this information could be digested

DOI: 10.4018/IJCWT.2020040101

into a format upon which predictions regarding insider threat cases could be made. Having said this, a data driven approach to insider threat mitigation is not a new idea, this is a field experiencing an increasing rate of publication. However, vanguard attempts still report more effective models than later cases where machine learning has been applied (Gheyas, 2016).

In machine learning/data mining projects, an imbalanced dataset is a dataset in which the number of observations belonging to one class is considerably lower than those belonging to other class/ classes. A predictive model employing conventional machine learning algorithms could be biased and inaccurate when being employed on such datasets. This is purely because machine learning algorithms are designed to improve accuracy by reducing the error in the network. Therefore, they do not consider the class distribution, class proportion, or balance of the classes in their classification process. A predictive machine learning model being bias or inaccurate can be predominant in scenarios where the minority class belongs to the malicious activities and the anomaly detection is extremely crucial. This includes scenarios such as: occasional fraudulent transactions in banks, irregular insider threats, rare disease identification, natural disaster such as earthquakes, and periodic malicious activities on critical infrastructures (e.g. infrequent attacks on nuclear power plants or water supply systems in a city). Given the importance of these scenarios, an inaccurate classification by a predictive machine learning model could cost thousands of lives or huge cost to individuals and/or organisations. There are several techniques to solve such class imbalance problems using various sampling/non-sampling mechanisms e.g. oversampling, undersealing and SMOTE as well as ensemble methods and cost-based techniques. However, the importance of an imbalanced dataset has not been clearly and adequately investigated in the literature particularly for machine learning-based solutions for insider threat detections.

Therefore, in this paper, our focus is on an extremely imbalanced dataset of insider threats where the number of events belonging to the malicious class is considerably lower than those belonging to the benign class. We use spread subsample (Weka. Class SpreadSubsample, 2018) as a popular balancing technique. The filter allows you to specify the maximum spread between the rarest class and the most common one. For example, given an imbalanced dataset, you may indicate that there should be a figure of 3:1 difference in class frequency. For this, the original dataset first fits in the memory then a random subsample of a dataset will be produced given the identified maximum spread between two classes. In this paper, we specify the maximum "spread" between the rarest class (i.e. malicious events) and the most common class (i.e. benign events) as "1" representing uniform distribution between two classes. This allows us to keep all the malicious events plus the equal number of the benign events selected randomly which results in having a uniform distribution of malicious and benign events.

In this paper, we raise the following specific research questions:

RQ1: Does balancing the dataset during the pre-processing phase improve metrics such as: Classification Accuracy (CA), Time taken to Build the model (TB), Time taken to Test the model (TT), True Positive (TP) rate, False Positive (FP) rate, Precision (P), Recall (R), and F-measure (F) in comparison with the same metrics but on an imbalanced dataset?

RQ2: What are the important parameters for each classifier that configuring them could have an impact on the classification results?

RQ3: Does changing these parameters with different values improve metrics such as: Classification Accuracy (CA), Time taken to Build the model (TB), Time taken to Test the model (TT), True Positive (TP) rate, False Positive (FP) rate, Precision (P), Recall (R), and F-measure (F) in comparison with running the classifiers with the default parameters?

Additionally, this paper provides a comprehensive explanation and investigation on the data pre-processing stage which is a crucial part of any data mining/ machine learning projects. For this, a clear step-by-step description is provided by the authors. Furthermore, we provided six comprehensive