ABSTRACT

MCoM (Mixup Contrastive Mixup) is a novel semi-supervised learning approach that introduces an innovative triplet mixup data augmentation technique to tackle the imbalanced data issue in tabular security datasets. In cybersecurity, tabular datasets are notorious for their severe data imbalances, where only a few labeled attack samples exist amidst a vast sea of mostly unlabeled benign data. While semi-supervised learning has been extensively explored in image and language domains, it remains relatively underutilized in security domains, particularly when dealing with tabular security data. This domain-specific challenge involves handling complex contextual information loss and data balance issues. Experimental results involving MCoM on collected security datasets demonstrate promising outcomes, achieving state-of-the-art performance when compared to alternative methods.

PROPOSED METHOD

MCoM trains a neural network on convex combinations of pairs of examples and their labels. Figure 1 illustrates our proposed MCoM framework, containing 4 parts: (1) triplet mixup data augmentation on the minority (vulnerable) class to address imbalance in the tabular security data set; (2) contrastive and feature reconstruction loss to train the encoder and the decoder; (3) pseudo-labeling of the subset of the unlabeled data using a label propagation technique; and (4) downstream tasks that train the predictor (e.g., MLP) with the fixed trained encoder.

The results demonstrate that our method achieved the best recall and TPR (66.67%) with just 0.1% labeled data, outperforming other methods, even those using all labeled data. Thus, MCoM exhibits superior performance across all metrics.

OBJECTIVES

The paper introduces MCoM, a novel semi-supervised machine learning approach designed for analyzing highly imbalanced tabular security datasets. MCoM consists of four key components: Triplet Mixup Data Augmentation, Contrastive and Feature Reconstruction Loss, Pseudo-labeling, and Downstream Tasks. When compared to six supervised methods and two state-of-the-art semi-supervised methods in tabular domains, it becomes evident that all methods perform poorly without the triplet mixup data augmentation, significant improvements are observed, with MCoM achieving the best recall and TPR at 66.67. Further research should explore the applicability of this technique to different datasets in various domains. Additionally, the authors plan to extend their method to graph datasets, including CCGFs, CPGs, and ASIs extracted from open-source applications, to identify software vulnerabilities at both source and binary levels. This approach aims to automate the labeling of graphs associated with each function and component, benefiting developers and experts.

REFERENCES


RESULTS

Table 1 summarizes the four different dimensions of features we studied associated with these vulnerabilities: structure-based, flow-based, binary-based, and pointer-based.

Table 2 displays experimental outcomes comparing our method with six supervised and two semi-supervised methods (labeled ratio 0.1). We divide the experiments into two sections: (1) without triplet mixup data augmentation (upper part) and (2) with triplet mixup data augmentation (lower part). In the upper section, without triplet mixup data augmentation, all methods predict only negative data and struggle with identifying positive (vulnerable) data, despite high accuracy and AUC on the negative samples. In contrast, using mixup generates more samples from the positive samples. SMOTE generates positive samples by adding small amounts to positive samples. However, our method leverages more information from positive samples by mixing-up triplet data points.

Table 3 compares different mixup strategies: pairwise, quadruplet, pairwise + original (mixing a pair of data points including the output of pairwise mixup), pairwise + triplet (mix a pair of data points followed by triplet mixup, including the output of pairwise mixup).

CONCLUSIONS

The paper introduces MCoM, a novel semi-supervised machine learning approach designed for analyzing highly imbalanced tabular security datasets. MCoM consists of four key components: Triplet Mixup Data Augmentation, Contrastive and Feature Reconstruction Loss, Pseudo-labeling, and Downstream Tasks. When compared to six supervised methods and two state-of-the-art semi-supervised methods in tabular domains, it becomes evident that all methods perform poorly without the triplet mixup data augmentation, significant improvements are observed, with MCoM achieving the best recall and TPR at 66.67. Further research should explore the applicability of this technique to different datasets in various domains. Additionally, the authors plan to extend their method to graph datasets, including CCGFs, CPGs, and ASIs extracted from open-source applications, to identify software vulnerabilities at both source and binary levels. This approach aims to automate the labeling of graphs associated with each function and component, benefiting developers and experts.