6ZC-02

機械学習を用いた攻撃検知のための オーバーサンプリング手法の一検討

松井 遼太朗 †1　ギリエ ルイス †2　和泉 諭 †3　水木 敬明 †1,†2　菅沼 拓夫 †1,†2
†1 東北大学大学院情報科学研究科　†2 東北大学サイバーサイエンスセンター
†3 仙台高専専門学校総合科学科

1　はじめに

近年，深刻化するサイバーセキュリティの脅威に対し，堅牢な攻撃検知・検出システムが求められている．現在，機械学習を用いた攻撃検知の研究が数多くされているが，機械学習の際に用いるデータセットには不均衡なデータが含まれていることが多く，効率的な学習が難しい問題がある．また，不均衡データの中でも多数派クラスと少数派クラスの境界部は誤分類が多く，高精度の検知が難しい問題がある．

データセットの不均衡問題を解消する手法として，多数派クラスのデータ群に合わせて少数派クラスのデータ群を増やすオーバーサンプリング手法がある．既存の研究の中で，多数派クラスと少数派クラスの境界部のデータ数を増加させるために効果的な手法は無く，未だ境界部の誤分類の問題を解決できていない．

そこで本研究では，多数派クラスと少数派クラスのデータの境界部のデータを集中的に増やすオーバーサンプリングアルゴリズムを提案する．

2　関連研究

2.1　SMOTE

SMOTE [1] は不均衡データによく使用されるオーバーサンプリング手法である．SMOTE は各少数派データから，ランダムに選ばれたデータとの間に新たな合成データを生成する手法である．しかしながら，SMOTE のアルゴリズムは，多数派クラスと少数派クラス分布割合が全く考慮されておらず，過度に少数派データを増やしてしまうため，検知精度を上げるための効果的なサンプリング手法とはいえない．

2.2　ADASYN

ADASYN [2] はデータの密度分布を利用して，少数派データを増やす手法である．多数派データの密度分布が高いときにはより多くの合成データを生成し，低いときには 0 又は少ない合成データを生成するアルゴリズムである．しかしながら，この手法では少数派データがまばらに分布しているデータセットに対しては効果が無く，データセットの次元が大きいと ADASYN のアルゴリズムが機能しないという問題がある．

2.3　Borderline-SMOTE

Borderline-SMOTE [3] は SMOTE のアルゴリズムを使用して境界部のデータ数を増やす手法である．まず，各少数派データに対して，k 個の最近傍の点を検知する．それらの点のうち，多数派クラスに属するデータが多いときは，ランダムな少数派データに対して SMOTE を実行するアルゴリズムである．しかしながら，この手法では少数派データがまばらに分布しているデータセットに対しては効果が無く，SMOTE のアルゴリズムが使われているため，境界部以外の箇所もデータが増えてしまうという問題がある．

3　オーバーサンプリング手法の提案

本章では提案するオーバーサンプリングアルゴリズムについて説明する．まず，少数派データのそれぞれに対して重みを計算する．そして最近傍の少数派データと多数派データに対して，別の割合で合成データを生成することで，Borderline-SMOTE よりも集中的に境界部のデータを増加させる．また，SMOTE のアルゴリズムのように少数派データが過度に増加する問題や ADASYN，Borderline-SMOTE のように意図しない箇所のデータが増加する問題を，少数派データ同士の距離による条件分岐を行うことで解決する．本研究の提案するアルゴリズムを Algorithm 1 に示す．

まず，少数派の各データに対して重み w を計算する．次に w が 0 の時は合成データを生成せず，1 の時は k 個分，少数派データに対して $0 \leq r \leq 1$ の範囲で合成データを生成をする．ただし，合成先の少数派データとの距離 dif が閾値以下の場合のみ合成データを生成し，閾値を超える時は合成データを生成しないという条件分岐を行う．

$0<w<1$ のときは，まず合成するデータ数 n を，最近傍の少数派データの w，n から計算する．そして n 個分，少数派データに対して $0 \leq r \leq 1$ の範囲で合成データを生成し，多数派データに対しても $0 \leq s \leq 1/2$ の範囲で合成データを生成する．

A Study of Oversampling Methods for Cyber Attack Detection using Machine Learning
Ryotaro MATSUI[†1], Guillen Luis[†2], Satoru IZUMI[†3], Takaaki MIZUKI[†1,†2], and Takuo SUGANUMA[†1,†2]
[†1]Graduate School of Information Sciences, Tohoku University
[†2]Cyberscience Center, Tohoku University [†3]National Institute of Technology, Sendai College

Algorithm 1 Proposed algorithm

Input: the whole trainning dataset T
Output: resampled dataset
for each minority class sample **do**
find its k nearest neighbors in T;
N_N is the number of neighbors from minority class;
$w_i \leftarrow N_N/k$ (i=1,2$\cdots N_N$)
if $w_i = 0$ **then**
do nothing
else if $w_i = 1$ **then**
$n_i \leftarrow k$
dif_j is the distance to the each nearest minority classes (j=1,2$\cdots n_i$)
if $dif_j \geq threshold$ **then**
Generate n_i new synthetic data for minority data in the range $0 \leq r \leq 1$
end if
else
n_j=w_j*n'_j/w'_j ($0<j'<j$)
Generate n_j new synthetic data for minority data in the range $0 \leq r \leq 1$, for majority data $0 \leq s \leq 1/2$
end if
end for

4 評価

提案するアルゴリズムが，既存の手法と比較して，どの程度境界部のデータが増加しているかを定量的に評価した．データセットは python で作成したものを用い，クラス数を 2，データ数を 10000，不均衡割合を 9:1 として作成した．実験に使用したデータを図 1 に示す．

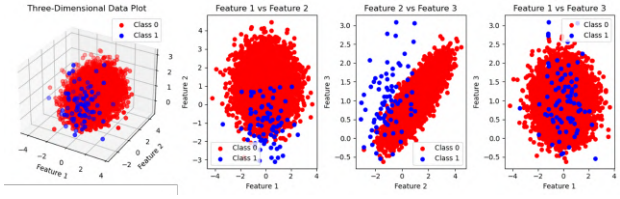

図 1: no sampling

提案手法，SMOTE，ADASYN，Borderline-SMOTE でオーバーサンプリングした結果をそれぞれ図 2，図 3，図 4，図 5 に示す．SMOTE ではランダムな箇所に新たなデータが生成されており，ADASYN では少数派データ周辺にデータが増加していることが分かる．Borderline-SMOTE では境界部を中心にデータが増加しているが，それ以外の箇所でもデータが生成されている，それに対し，提案手法では境界部のみデータが増えており，それ以外の箇所では増えておらず，境界部のデータが集中的に増えていることが分かる．

5 おわりに

本稿では，データセットの不均衡問題を解決するために，多数派クラスと少数派クラスの境界部にアプローチしたオーバーサンプリング手法を提案した．提案手法では既存手法に比べ，境界部のデータのみを増加させることができ，精度の高い機械学習モデルを構築できると考えられる．今後の課題としては，提案アルゴリズムの定量的な評価や，より多くの種類の機械学習モデルで実験，提案手法を実際の異常検知のデータセットに適用した時の効果の有無を検証する必要がある．

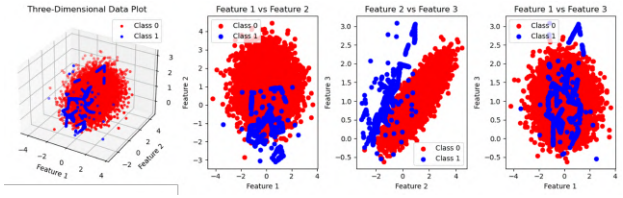

図 2: Proposal

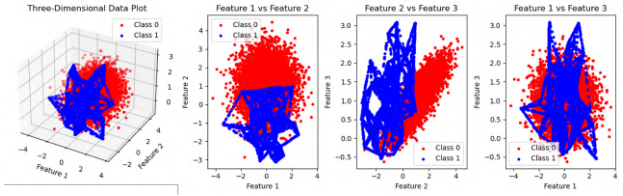

図 3: SMOTE

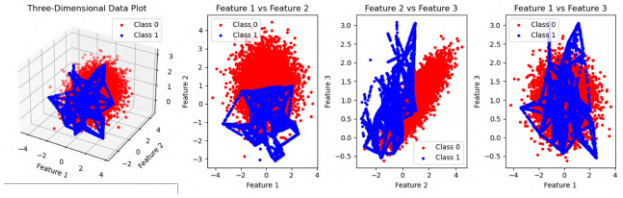

図 4: ADASYN

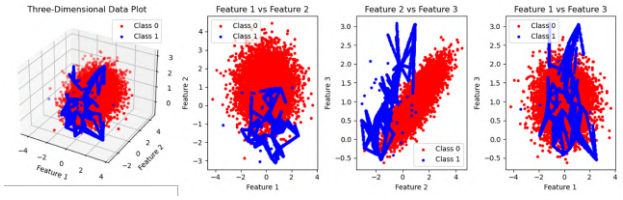

図 5: Borderline-SMOTE

参考文献

[1] Chawla, N. V. et al.: SMOTE: synthetic minority over-sampling technique, *Journal of artificial intelligence research*, Vol. 16, pp. 321–357 (2002).
[2] He, H. et al.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328 (2008).
[3] Han, H. et al.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, *International conference on intelligent computing*, pp. 878–887 (2005).