令和５年度卒業研究発表会（日本大学工学部情報工学科）
C-2-22

スケーリング処理を用いたネットワークトラフィックの異常検知

Scaling in network traffic anomaly detection

20206181　　山内　優輝　　［和泉研究室］

1.　まえがき

近年,サイバー攻撃に関連する通信数は増加傾向にあり,不正なネットワークトラフィックを検出することはセキュリティ対策を行う上で重要となっている.ネットワークトラフィックのデータには,他の観測値に比べて異常に大きい値の外れ値が存在する場合がある.機械学習を用いたネットワークトラフィックの異常検知ではこの外れ値が検知率に大きく関わっており,適切なスケーリング処理を行う必要がある.

そこで本研究では,サポートベクターマシンを用いてネットワークトラフィックの異常検知を行い,その際,スケーリング処理が検知率にどのような影響を与えるか検証していく.

2.　サポートベクターマシン[1]

サポートベクターマシン（以下 SVM）とは，教師あり機械学習アルゴリズムの１つであり，２つのクラスのデータ群を分割するような境界線や超平面を決定することで分類や回帰を行うアルゴリズムである.境界線から最も近いデータ(サポートベクトル)と境界線との距離(マージン)を最大化することで，分類性能の向上を行う.

本研究では線形分離不可能なデータを前提として、誤判別を許容するソフトーマージンの SVM を用いる.

3.　提案手法

本研究では，ネットワークトラフィックのデータを SVM に学習させ異常検知を行う.学習を行う際に検知精度の向上を目的として，データに対して正規化，標準化，逆数変換のそれぞれのスケーリング処理を行う.

スケーリング処理によって,適切な学習を行い検知精度の向上を期待する.

4.　スケーリング処理

4．1　正規化

データの各特徴の値の範囲を最小値０，最大値１に変換する．特徴の各値をx^i，特徴の最小値をx_{min}，特徴の最大値をx_{max}とした際，以下の式となる.

$$x^i_{\mathrm{norm}} = \frac{x^i - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

4．2　標準化

データの各特徴の値を平均０，分散１に変換する．特徴の各値をx^i，特徴の平均値をu，特徴の標準偏差をσとした際，以下の式となる.

$$x^i_{std} = \frac{x^i - u}{\sigma} \qquad (2)$$

4．3　逆数変換

データの値を分数の形にし,分母と分子を入れ替える処理を行う．特徴の各値をx^iとした際，以下の式となる．x^iの値が０の場合,計算できないため，ε＝０．００１とする.

$$x_{reci} = \frac{1}{x^i + \varepsilon} \qquad (3)$$

5.　実験

5．1　実験方法

本実験では実験データとして NSL-KDD データセット[2]を使用する.このデータセットは通信データをセッション単位で加工したものであり，４１種類の特徴[3]がある．特徴の中にはカテゴリ変数が含まれているため One-hot encoding を用いてダミー変数を作成する前処理を行う.

また.本実験ではデータの範囲が他に比べて著しく大きかった２つの特徴（図１，図２，図３）を外れ値を含んでいる特徴と仮定し，実験を行う.

実験データは学習データとして１２５９７３個のセッション,テストデータとして２２５４４個のセッションを用いる．この実験データに正規化，標準化，逆数変換のスケーリング処理をそれぞれ用いて実験を行い,検知率の比較を行う.

1. NSL-KDD データセットの学習データ，テストデータにそれぞれ One-hot encoding を用いたダミー変数処理を行う.
2. 各データセットに対して，正規化，標準化，逆数変換のスケーリング処理をそれぞれ行い学習データを作成する．また，比較のため，外れ値を含む特徴のみスケーリング処理を行う場合と,外れ値を含まない特徴のみスケーリング処理を行う場合の学習データも作成する.
3. スケーリング処理後の学習データを SVM に学習させる．また，比較のため，スケーリング処理を利用しない学習データを SVM に学習させたモデルも作成する.
4. それぞれモデルに対して同様の処理を行ったテストデータを入力することで正常か異常か推定し,その性能を比較する.

5．2 One-hot encoding

One-hot encoding は，カテゴリ変数の数に応じて特徴種別を増やし，そのカテゴリが含まれていた場合１，含まれていない場合０に変換する（ダミー変数）処理である.

5．3 実験結果

実験によって得られた結果を表１に示す.正答率は学習データから学習した SVM のモデルに，テストデータを入力した際の正常と異常の分類精度を示す.

表 1　各処理の正答率

処理	正答率
スケーリング処理なし	43.1%
すべての特徴を正規化	74.6%
外れ値を含む特徴のみ正規化	68.9%
外れ値を含まない特徴のみ正規化	43.1%
すべての特徴を標準化	79.3%
外れ値を含む特徴のみ標準化	69.0%
外れ値を含まない特徴のみ標準化	43.1%
すべての特徴を逆数変換	78.5%
外れ値を含む特徴のみ逆数変換	78.0%
外れ値を含まない特徴のみ逆数変換	43.1%

6.　考察

実験結果から，外れ値を含む特徴に処理を行わなかった際に，正答率があまり変わらなかったことから，外れ値を含む特徴が分類精度に大きく関わっていることが確認できる．

本実験では，すべての特徴を標準化した際の正答率が一番高い結果となった．標準化を行うことですべての特徴の値の単位が等しくなる．よって，SVM へ学習させる際，データの単位が等しくなるようにスケーリング処理を行うことが適切な処理だったと考えられる．

また，３つのスケーリングの処理のうち，正規化が一番正答率の低い結果となった．これは１～０に正規化を行う際に，外れ値によって０付近に値が偏り，他のスケーリング処理比べて適切に学習が行われなかったためと考えられる．

一方で，逆数変換ではすべての特徴を変換したときと，外れ値を含む特徴のみを変換したときとで正答率があまり変化しなかった．これは，逆数変換が外れ値のような大きな値に対して効果的な処理だが，外れ値を含まない散らばりの少ない特徴の場合，影響の少ない処理であると考えられる．

7.　むすび

本研究では，機械学習を用いたネットワークトラフィックの異常検知の際の適切なスケーリング処理の検討を行った．外れ値を含む特徴に対してスケーリング処理を行った際，すべての処理で分類精度が向上し，特にすべての特徴に対して標準化を行った際の分類精度が一番高い結果となった．

今後の課題として，精度の向上があげられるが，分類精度へ影響の大きい特徴の特定，また，その特徴に対してどのような処理が適切なのか検証していく必要があると考えられる．

参考文献

[1][Python]サポートベクトルマシン（SVM）の理論と実装を徹底解説してみた
https://qiita.com/renesisu727/items/964005bd29aa680ad82d
[2] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set”, Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.
[3] Selecting Optimal Subset of Features for Intrusion Detection Systems - Scientific Figure on ResearchGate. Available from: https://www. researchgate. net/figure/The-41-features-of-NSL-KDD-dataset_tbl1_287302551 [accessed 23 December, 2023]

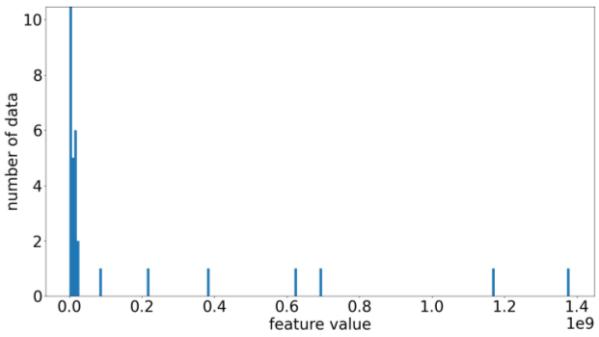

図１　外れ値を含んだ特徴のヒストグラム

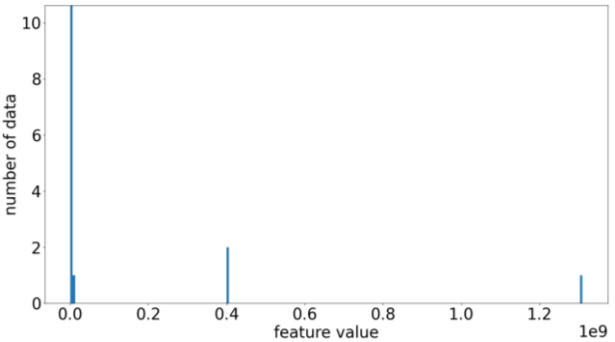

図２　外れ値を含んだ特徴のヒストグラム

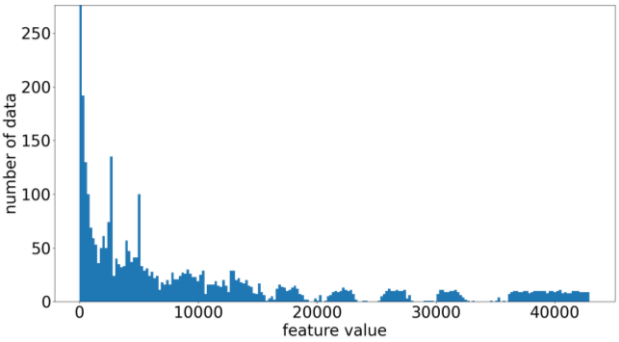

図３　外れ値を含んでいない特徴のヒストグラム