異なる機械学習アルゴリズムと 4 つの特徴選択法による DDoS 攻撃検出 の パフォーマンス比較

秋山 仁志 †1 フォン ヤオカイ †2 櫻井 幸一 †3
九州大学大学院システム情報科学府 †1 九州大学大学院システム情報科学研究院 †2,3

概要：DDoS 攻撃検出手法としては機械学習を使った手法が最も多く存在している。本研究では、RandamForest、LightGBM、Lasso、ElasticNet について特徴選択数、分類精度、プログラムの実行時間の関係を比較し考察した。

キーワード: 侵入検出・検知、不正・異常検出、分類学習

Performance comparison of DDoS attack detection with different machine learning algorithms and four feature selection methods

Hitoshi Akiyama†1 Yaokai Feng†2 Koichi Sakurai†3
Kyushu University Graduate School of Information Science and Electrical Engineering†1†2.3

Abstract: : Most DDoS attack detection methods use machine learning. In this study, we compared RandamForest, LightGBM, Lasso, and ElasticNet by comparing the relationship between the number of feature selections, classification accuracy, and program execution time.

Keywords: Intrusion detection・detection, Fraud ・abnormality detection, Classification learning

1. はじめに

DDoS 攻撃は、企業および個人全体で最も持続的かつ重大な脅威の 1 つとなっている。攻撃者は、さまざまな隠蔽技術を利用し、乗っ取った複数のコンピューター（ボットネット）が検出されないようにする。したがって、DDoS 攻撃の脅威に対抗することは困難な作業であり、ボットネットの検出のためにさまざまな研究が提案されている。また、近年モノのインターネット機器(IoT)に関連したネットワーク攻撃も全体の半数ほどに増えてきていると言われている。その中で、代表的なマルウェア Mirai は、IoT を用いてボットネットを形成し DDoS 攻撃を行い、甚大な被害を与えた。

ボットネットの検出手法としては機械学習を使った手法が最も多く存在している。機械学習を行う際には、前段階としてどの特徴量を使うのかという選択（特徴選択）が重要である。機械学習を用いた特徴選択法における代表的なものは、決定木を弱学習機としてのアンサンブル学習、ランダムフォレスト(RF)や、ロジスティック回帰の一つである Lasso 回帰(Lasso)などである。LightGBM(GBM)と Elastic Net(EN)はそれぞれ既存の特徴選択法として代表的な RF と Lasso を発展させたものである。GBM は RF と同様に決定木を弱学習機とするアンサンブル学習でブースティングを使用したもの、EN は Lasso と同様にロジスティック回帰の一つであり、Lasso の正則化項に修正を加えたものである。また、RF と Lasso による分類性能の比較を行っている研究がいくつか存在している。一方で、RF と Lasso の発展させた GBM と EN についての比較はいまだ行われていない。

本研究では、公開されている 2 つのデータセットを用いて RF、GBM、Lasso、EN の 4 つの特徴選択数、分類精度、プログラムの実行時間の関係を比較し、GBM と EN に着目して考察した。1 つ目のデータセットは特徴数 55 個、データ数 68264 個のもの。2 つ目のデータセットは特徴数 41 個、データ数 1782 個のものである。その結果、分類精度においては、GBM は特徴数を 10 個まで減らしても 99%と高い分類精度を維持したのに対し、EN の分類精度は全ての特徴を使用した状態で 72%、使用する特徴数を少なくすると 41%まで分類精度が落ちた。実行時間については、ほとんどの場合で GBM が優れており、最大で EN に比べて 74 倍の速さであったが、データセット 1 において使用する特徴数 17 個以下の範囲では、最大で 9.7 倍 EN のほうが速いという結果であった。これらから、効果的に DDoS 攻撃を検知

1 九州大学大学院システム情報科学府
Kyushu University Graduate School of Information Science and Electrical Engineering

する手法として RF よりも GBM の方が優れていることを明らかにした。

2. ネットワーク攻撃

インターネットはここ数十年で大きく進化した。また、インターネットの使用を急速に増加させ、同時にユーザー数を大幅に増加させる Web アプリケーションの種類が発明された。1994 年のインターネットユーザー数は 2,500 万人だったが、2016 年には 34.24 百万人となり、136 倍に増加した。 2020 年の時点で、ユーザーの数は 4 億 4,590 万人で、2016 年から 1.3 倍増加しており、世界の人々の約 40%がインターネットを使用している[1]。インターネットは私たちの生活に必要なツールになっており、インターネットサービスの可用性を確保することが重要である。ただし、インターネットの世界には依然として多くのサイバー攻撃がある。これらの攻撃者の一部は、マルウェアの拡散、電子メールフィッシング、アカウントハイジャックなどのインターネットユーザー向けの攻撃方法を使用する場合がある。攻撃者は、偽のメールやホームページを用意して、ユーザーをだましてアカウントにログインさせ、ユーザーのコンピューターはマルウェアのダウンロードを自動的に開始し、ウイルスに感染する。

サーバーから不正に情報を取得する以外の別のサイバー攻撃は、サーバーサービスをクラッシュさせようとする DDoS 攻撃である。 DDoS 攻撃は、実際には DoS 攻撃の一種であり、サービス拒否攻撃を意味する。DDoS 攻撃を行う最も効率的な方法は、HTTP 要求パケットなどの正当なパッケージを大量に送信し、サーバーがすべての要求に応答するのを困難にすることである。

表 1 サイバー攻撃における調査（2015~2017）

year	malware	Account hijacking	Targeted attack	DDoS attack	SQL injection
2015	6.4%	8.8%	10.5%	9.7%	17.5%
2016	7.9%	15.0%	11.5%	11.2%	8.3%
2017	29.8%	15.6%	15.2%	4.2%	0.8%

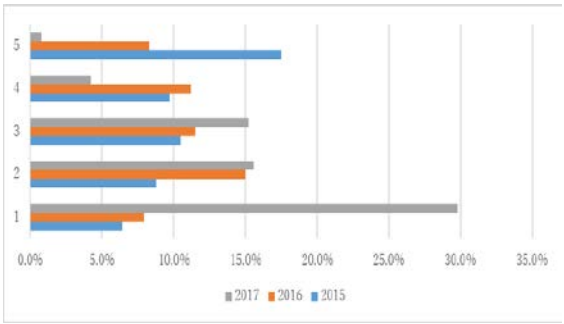

図 1 サイバー攻撃における調査（2015~2017）

表 1 と図 1 は、近年のサイバー攻撃の割合に関する統計を示している[2]。 1 はマルウェア、2 はアカウントハイジャック、3 は標的型攻撃、4 は DDoS 攻撃、5 は SQL インジェクションである。この表は、DDoS 攻撃が 2016 年にはサイバー攻撃の 11.3%を占めていたが、2017 年には 4.2%に低下したことを示している。

3. DDoS 攻撃とボットネット

DDoS 攻撃は、サーバーのネットワークリソースを使い果たしてユーザーにサービスを提供する分散型サービス拒否攻撃として定義されている[3]。DDoS 攻撃はサーバーを標的としており、ユーザーは攻撃しない。ただし、DDoS 攻撃とマルウェア拡散攻撃の間には強いつながりがある。DDoS 攻撃を実行するには、膨大な数のマシンを必要とする巨大なリクエストパケットを生成する必要がある。これらのマシンを維持する効果的な方法は、マルウェアを使用して通常のユーザーのコンピューターを制御することである。実際、侵害されたコンピューターは、ボットネットと呼ばれる DDoS 攻撃の構造の 1 つになる。図 2 は、攻撃者、C&C サーバー、ボットネット、標的サーバーを含む DDoS 攻撃の一般的な構造を示している。ボットネットはハイジャックされたコンピューターのネットワークであり、ボットネットと攻撃者の間の接続は、コマンドアンドコントロールサーバー（C&C サーバー）と呼ばれている。

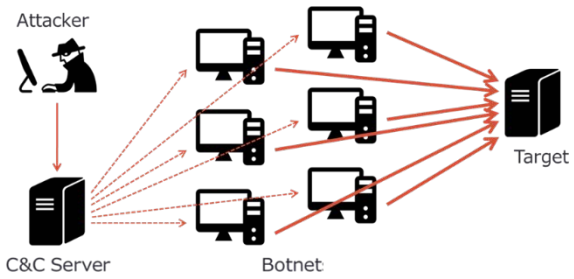

図 2 DDoS 攻撃の構造

4. 関連研究

A.ペクタシュ、T。アカマンらは、ボットネット検出のための 3 つの機能選択法、3 つの機械学習を使用して、合計 9 通りのボットネット検出のパフォーマンスを比較した[4]。使用された特徴選択方法は、Lasso、再帰的特徴除去（RFE）、およびランダムフォレストであった。

ランダムフォレストアルゴリズムは、分類学習に使用される一般的な機械学習であり、各特徴の重要度を分類学習と同時に計算できるため、特徴選択にも使用される。Lasso は、回帰分析中の過剰学習を防ぐために、特徴選択の最小

二乗コスト関数に正則化項を追加することで罰則を設ける。Lasso は、データにノイズがなく、フィーチャ間に相関がない場合にうまく機能するという特徴がある。再帰的特徴除去とは、モデルを繰り返し構築し、パフォーマンスの低い機能を削除することにより、機能を選択する方法である。実験には ISOT データセットが使用された。このデータセットは、ボットからの悪意のあるデータセットと、Web サーフィン、ファイル共有などを含む通常のデータセットの組み合わせである。

実験結果を表 3.6 に示す。 縦軸は使用される特徴選択方法を示し、横軸は使用される分類学習アルゴリズムを示す。その結果、ランダムフォレストによって実行された特徴選択と異常分類は、99%の精度で 9 パターンの中で最高のパフォーマンスを示した。 また、最高のパフォーマンスを示すときに選択された 9 つの機能は、uration、numBytesRcvd、minPktSz、maxPktSz、avePktSize、stdPktSize、pktAsm、bytAsm、および tcpMinWinSz であった。

表 2　異なる特徴選択、機械学習によるパフォーマンス比較結果

	Machine Learning Algorithms											
	Random Forest				Logistic Regression				SVM			
Feature Selection	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Tree Based	0.99	0.99	0.99	0.995	0.72	0.71	0.61	0.710	0.88	0.87	0.87	0.901
RFE	0.94	0.95	0.94	0.943	0.86	0.82	0.84	0.849	0.90	0.92	0.91	0.927
Lasso	0.92	0.92	0.92	0.936	0.85	0.84	0.84	0.891	0.88	0.89	0.88	0.912

5. 研究の提案と実験

5.1 研究の提案

以前の研究では、ランダムフォレストと Lasso、再帰的特徴除去などの比較実験が行われており、ランダムフォレストにより特徴選択法を行い、分類学習をしたものが最も効果的であった。 ただし、ランダムフォレストを進化させた LightGBM と Lasso を進化させた Elastic Net の比較実験は行われていない。本研究では、ランダムフォレスト、LightGBM、Lasso、Elastic Net の 4 つの方法を使用し、分類性能を比較、結果を検討する。

5.2 ランダムフォレストと LightGBM

ランダムフォレストは、アンサンブル学習の一種であり、与えられたトレーニングデータから複数の決定木を作成することでモデルを生成することができる[5]。 予測を行うとき、各決定木から出力される多数の結果が出力されることはない。 ランダムフォレストは、図 3 に示すように、アンサンブル学習でバギングと呼ばれるトレーニングを使用する。ランダムに抽出されたデータセットが各決定ツリーに使用される。

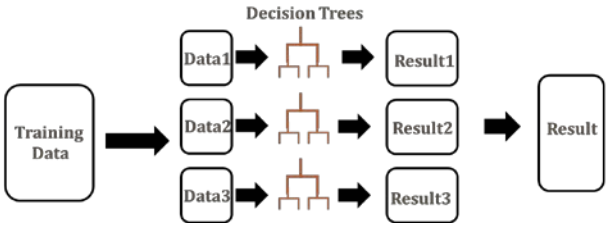

図 3 ランダムフォレストの構造

LightGBM は、意思決定ツリーを使用したアンサンブル学習であり、これはランダムフォレストと同じである[6]。ランダムフォレストとの違いは、図 4 に示すように、ランダムフォレストは並列にトレーニングコースを使用し、LightGBM はブースティングにトレーニングコースを使用することである。その結果、決定木 2 は、決定木 1 の弱点を補うように決定木を作成するため、バギングアルゴリズムよりも高い分類精度のパフォーマンスを発揮できる。ただし、バギングアルゴリズムは並行して処理できるため、実行時に高いパフォーマンスを発揮できる。

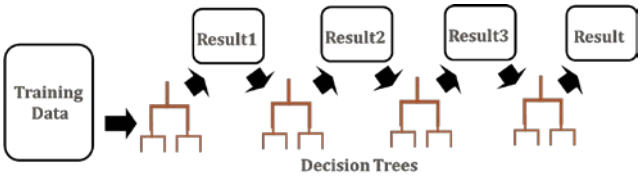

図 4 LightGBM の構造

5.3 Lasso と Elastic Net

L1 正則化は、正則化された線形回帰の 1 つである。正則化は、過剰学習を防ぎ、汎用性を高めるための手法であり、モデルが複雑になりすぎないように正則化の用語をモデルに追加することにより、モデルの形状を調整するものである。L1 正則化(Lasso)は、罰則項として $\Sigma|\beta|$ を用いる。Lasso には、独立変数間の相関が高いグループから 1 つの独立変数がモデルに対してランダムに選択される。また、β の値が 0 になると、未使用の特徴が生成され、特徴選択を行うことができる[7]。

Elastic Net は Lasso の改良版であり、機能間の高い相関を考慮できる。[8]罰則として、L1 正則化と L2 正則化の両方が使用される。L1 正則化の特徴選択を実行できるという特徴と、L2 正則化は強い相関特徴をモデルに組み込むことができるという特徴をどちらも持っている特徴選択法である[9]。

5.4 実験手順

本研究では、55 個の特徴を持つデータセットと 48 個の特徴を持つデータセットを使用した。60%がトレーニングデータに、40%がテストデータに分割される。ランダムフォレストと LightGBM を使用して、以下の手順を実行する。

まず、すべての機能を使用して分類学習が実行され、分類の精度と実行時間を記録する。同時に、機能の重要度が計算される。次に、重要性の低い機能からいくつかの機能を削除して、新しいデータセットを作成する。次に、この新しいデータセットを使用して分類学習が実行され、分類精度と実行時間を記録すると同時に、特徴の重要度が計算される。これらの 3 つのステップを特徴の数が十分に小さくなるまで繰り返す。

Lasso および Elastic Net の場合、以下の手順に従う。 まず、モデルの係数に対する制約である 0.0001 のアルファ値で分類精度と実行時間を記録する。 次に、使用する機能の数を減らすためにアルファの値を徐々に増やし、分類の精度と実行時間を記録する。 これは、機能が十分に小さくなるまで実行される。 アルファ値を大きくすると、ペナルティが増加し、使用される機能の数が減少する。このように、2 つのデータセットと、4 つの特徴選択方法によって 8 つの実験結果を取得した。

6. 結果と結論

図 5 に、データセット 1 における特徴数と分類精度による実験結果を示し、図 6 に、データセット 2 における特徴数と分類精度による実験結果を示す。図 7 に、データセット 1 における特徴数と実行時間による実験結果を示し、図 8 に、データセット 2 における特徴数と実行時間による実験結果を示す。

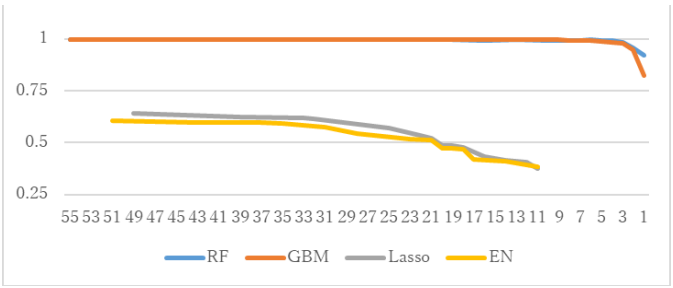

図 5 分類精度と特徴数によるパフォーマンス比較
（データセット 1）

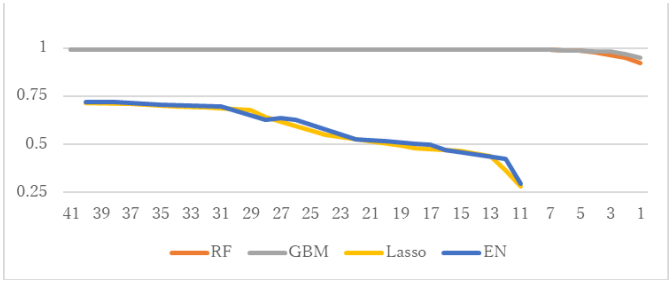

図 6　分類精度と特徴数によるパフォーマンス比較
（データセット 2）

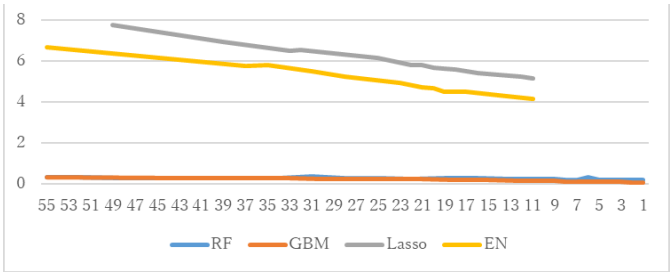

図 7　実行時間と特徴数によるパフォーマンス比較
（データセット 1）

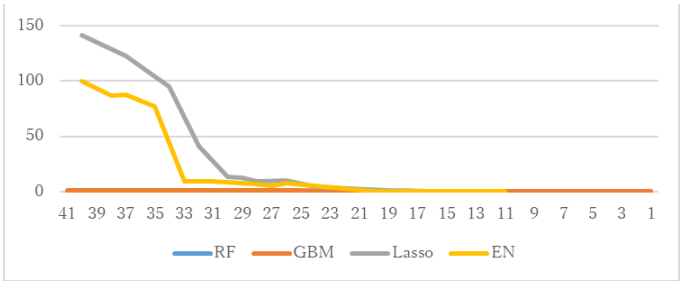

図 8　実行時間と特徴数によるパフォーマンス比較
（データセット 2）

本研究では、公開されている 2 つのデータセットを用いて RF、GBM、Lasso、EN の 4 つの特徴選択数、分類精度、プログラムの実行時間の関係を比較し、GBM と EN に着目して考察した。1 つ目のデータセットは特徴数 55 個、データ数 68264 個のもの。2 つ目のデータセットは特徴数 41 個、データ数 1782 個のものである。その結果、分類精度においては、GBM は特徴数を 10 個まで減らしても 99%と高い分類精度を維持したのに対し、EN の分類精度は全ての特徴を使用した状態で 72%、使用する特徴数を少なくすると 41%まで分類精度が落ちた。実行時間については、ほとんどの場合で GBM が優れており、最大で EN に比べて 74 倍の速さであったが、データセット 1 において使用する特徴数 17 個以下の範囲では、最大で 9.7 倍 EN のほうが速いという結果であった。これらから、効果的に DDoS 攻撃を検知する手法として RF よりも GBM の方が優れていることを明らかにした。

謝辞

本研究は、国立研究開発法人科学技術振興機構（JST）戦略的国際共同研究プログラム（SICORP）および JSPS 科研費 JP18K11295 と 17K00187 の助成を受けたものである。

参考文献

[1]"Internet Users"Internet live stats, "Internet Users",https://www.internetlivestats.com/internet-users/ accessed:29 Jan 2020.

[2]HACKMAGEDDON,"2017 Cyber Attacks Statistics", https://www.hackmageddon.com/2018/01/17/2017-cyber-attacks-statistics/ accessed:29 Jan 2020.

[3]Tomer Shani, "This DDoS Attack Unleashed the Most Pokets Per Second Ever. Here's Why That's Important", https://www.imperva.com/blog/this-ddos-attack-unleashed-the-most-packets-per-second-ever-heres-why-thats-important/accessed:29 Jan 2020.

[4] A. Pektaş, T. Acarman(2017) “EFFECTIVE FEATURE SELECTION FOR BOTNET DETECTION BASED ON NETWORK FLOW ANALYSIS”International Conference Automatics and Informatics'2017.

[5]Granitto PM, Furlanello C, Biasioli F, Gasperi F. “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products”, Chemometrics and Intelligent Laboratory Systems. 2006 Sep 15;83(2):83-90.

[6]Guolin Ke , Qi Meng , et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”,31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[7]Zhou Y, Jin R, Hoi S. “Exclusive lasso for multi-task feature selection”, InInternational conference on artificial intelligence and statistics 2010 (pp. 988-995).

[8]Wataru Sakamoto, Fumiaki Takahashi, Masahiro Takeuchi, “Study on variable selection method for multi-dimensional data by logistic regression model using regularization”, Laboratory of Mathematical Analysis, p32-35 1703 2010.

[9]Hitoshi Akiyama.”Comparison of performance of DDoS attack detection among different machine learning algorithms with feature selection methods”.Feb.2020