侵入プロセスの特徴を利用した侵入検知のためのソフトウェアアーキテクチャ

―相関ルールマイニングを用いた U2R 攻撃検知の精度向上に向けて―

2020SE033 松下侑加　2020SE084 吉田哲

指導教員：沢田篤史

1　はじめに

ネットワーク攻撃に対する侵入検知システムに相関ルールマイニングを活用した機械学習技術が広く利用されている．分析者は相関ルールを見てデータ間の関係性を知ることができるので，侵入検知システムの結果の理解性が高いという利点を持つ．

Brahmi ら [2] による研究では，U2R 攻撃に対する検知率が低いと報告されている．U2R 攻撃とは様々な方法でコンピュータ内に侵入し，時間差で管理者権限を奪う複雑なプロセスの攻撃である．

本研究の目的は複雑なプロセスを持つ攻撃の侵入プロセスを考慮した侵入検知システムの設計である．U2R 攻撃のような複雑な侵入プロセスを持つ攻撃の特徴を捉え，侵入検知の精度を向上させるためのシステム設計について検討する．

研究目的を達成するための技術課題として「複雑な侵入プロセスを捉えることができるソフトウェア構造を明らかにすること」と「ソフトウェアアーキテクチャの有用性の評価」の 2 点を設定する．

これらの技術課題に対して本研究では「プロセスの特徴に応じて役割分担を行う 2 段階アーキテクチャの提案」と「NSL-KDD データセットを用いた評価」を試みる．

本研究では 2 段階アーキテクチャを提案する．相関ルール抽出時の 1 段階目は侵入プロセスの特徴を考慮してデータを抽出し，2 段階目は抽出したデータから相関ルールを抽出する．

テスト時の 1 段階目では抽出された相関ルールを用いて U2R 攻撃の兆候とみられるデータを検出し，2 段階目に検出されたデータから時間に関する一定の閾値分後のデータをファイルに出力する．データ内に U2R ラベルのデータがある場合 U2R 攻撃の兆候を検知できたといえる．

本研究では 4 つの実験で U2R 攻撃の兆候であるデータの相関ルールの抽出を行う．相関ルールに入力する 2 つのデータ抽出方法と 2 つの属性選択方法を組み合わせた 4 つの実験を行う．4 つの実験結果を示し，2 段階アーキテクチャの有用性を考察する．

2　相関ルールマイニングを利用した侵入検知システムの問題

2.1　相関ルールマイニング

相関ルールマイニングは，データセットの中で多頻度で共起している項目を発見し，ルールを生成するマイニング方法である [5]．侵入検知に使われる際は，訓練時にネットワーク通信パケットのデータセットを入力することでデータ間の属性の関係性を調査し相関ルールを抽出する．

抽出されたルールを用いて侵入検知を行うことの利点は次の 2 点がある．

1 検知結果の理解性が高い

2 データ分析者の知識や経験を反映することができる

1 つ目は相関ルールによってデータ間の属性の関係性を知ることができるからである．2 つ目は抽出されたルールを侵入検知システムに反映する際に，データ分析者の判断で採用するルールを決定できるからである．

2.2　相関ルールマイニングを用いた侵入検知システムの関連研究

Brahmi ら [2] による研究では，OLAP（On Line Analytical Processing）と相関ルールマイニング技術を統合した OMC-IDS（OLAP Mining and Classification-based IDS）が紹介されている．彼らの研究では入力するデータでデータキューブを作成し相関ルールマイニングを実行する．

Brahmi らはデータキューブの次元数が高いほど，DoS 攻撃，U2R 攻撃，R2L 攻撃の 3 種類の攻撃検知率が向上する傾向があることを明らかにした．1 次元から 6 次元のデータキューブを用いた実験により，DoS，R2L，U2R 攻撃は 6 次元の時にそれぞれ最高検知率に達成している．また Probe 攻撃は 5 次元の時に最高検知率に達している．

2.3　相関ルールマイニングを用いた U2R 攻撃検知の課題

相関ルールマイニングには侵入検知結果の理解性が高いという利点があるが，Brahmi ら [2] の研究結果では U2R 攻撃の検知率が比較的低いことが示されている．これは相関ルールマイニングでは複雑な侵入プロセスを持つ U2R 攻撃の侵入の特徴を捉えることができていないことが原因である．

U2R 攻撃は，はじめに様々な手口でコンピュータ内に侵入し，最終的に管理者権限を奪い取る攻撃の総称である．U2R 攻撃はコンピュータ内に侵入した後に時間差で管理者権限を奪う．最終的な攻撃の際にはコンピュータ内から攻撃を試みるので攻撃時の通信パケットデータを監視することによる U2R 攻撃の検知は難しい．

3　研究目的と技術課題

3.1　侵入検知システム作成の目的

本研究の目的は複雑なプロセスを持つ攻撃の侵入プロセスを考慮した侵入検知システムの設計である．本研究では複雑な侵入プロセスを持つ攻撃として U2R 攻撃に焦点を当てる．U2R 攻撃を検知するには U2R 攻撃の侵入時のパケットデータを検知することが重要である．相関ルールを抽出するためのパケットデータの抽出時に侵入プロセスの特徴を捉えることで侵入検知システムの精度向上を目指す．

相関ルールマイニングを用いる理由として「検知結果の理解性の高さ」と「データ分析者の知識や経験を反映できること」の 2 点がある．相関ルールマイニングは学習結果として相関ルールが抽出されデータ間で属性の関係性を明確にできるので，侵入検知システムの検知結果の理解性が高い．抽出されたルールは可読性があることから分析者の経験や知識を踏まえて侵入検知を行うことも可能である．

理解性が高くても検知率が低ければルールの信頼性が損なわれてしまう．相関ルールマイニングを用いた先行研究では U2R 攻撃の検知率が比較的低いことが報告されている．

3.2　侵入プロセスの特徴を考慮した侵入検知システムの技術課題

研究目的を達成するための技術課題は次の 2 点である．

- 複雑な侵入プロセスを捉えることができるソフトウェア構造を明らかにすること
- ソフトウェアアーキテクチャの有用性の評価

1 つ目は U2R 攻撃のような複雑な侵入プロセスを持つ攻撃の特徴を捉えることができるソフトウェアアーキテクチャの設計である．U2R 攻撃は 1 段階目にコンピュータへ侵入し，2 段階目に管理者権限を奪うという攻撃であり，U2R 攻撃の特徴として 1 段階目に攻撃をせず侵入を試みることである．この特徴を捉えることができるアーキテクチャを設計することで，複雑な侵入プロセスを持つ攻撃の 1 つである U2R 攻撃の検知率の精度を向上させることができる．

2 つ目は提案したソフトウェアアーキテクチャの有用性の評価である．設計したアーキテクチャを実装することにより本研究の目的である複雑な侵入プロセスを持つ攻撃の検知率の向上が達成されているか評価する．

4　2 段階目アーキテクチャを用いた侵入検知システムの設計

4.1　2 段階アーキテクチャ

複雑な侵入プロセスの特徴を捉えることができるアーキテクチャとして，プロセスの特徴に応じて役割分担を行う 2 段階アーキテクチャを提案する．提案する 2 段階アーキテクチャは相関ルール抽出時とテスト時の構造が 2 段階となる．

4.2　相関ルール抽出時の 2 段階アーキテクチャによる相関ルールマイニングの抽出

2 段階アーキテクチャによる相関ルールマイニングを抽出する方法の概要を図 1 に示す．

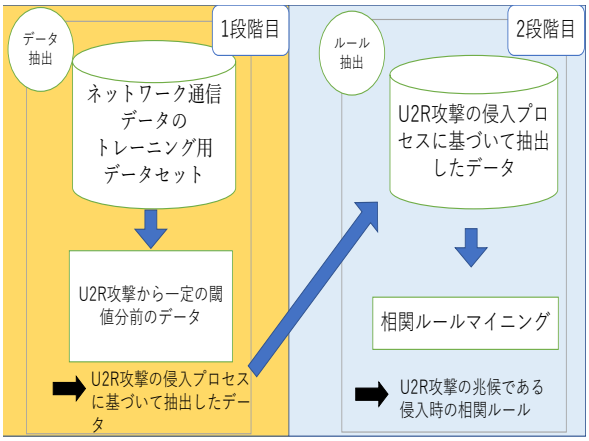

図 1　相関ルール抽出時の 2 段階アーキテクチャ

2 段階アーキテクチャの 1 段階目では，トレーニング用データセットの U2R ラベルのあるデータから一定の閾値分前のデータを抽出することで兆候と見られるデータを抽出する．一定の閾値分前のデータを抽出する理由は，U2R 攻撃は管理者権限を奪う前にコンピュータ内への侵入をするプロセスを持つからである．

2 段階目では，1 段階目で抽出されたデータを相関ルールマイニングに入力しデータ間の相関ルールを抽出する．1 段階目で U2R 攻撃の侵入プロセスに着目しデータを抽出したので，2 段階目で抽出される相関ルールは U2R 攻撃の兆候である侵入時の相関ルールが抽出される．

4.3　2 段階アーキテクチャによる相関ルールのテスト

2 段階アーキテクチャによる相関ルールのテストの概要を図 2 に示す．

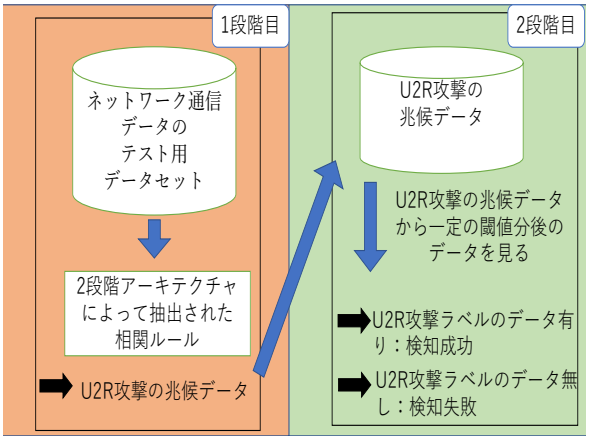

図 2　テスト時の 2 段階アーキテクチャ

テスト時の 2 段階アーキテクチャの 1 段階目ではテス

ト用データセットから相関ルール抽出時に抽出された相関ルールに従ってデータを抽出する．使用する相関ルールは U2R 攻撃の兆候と見られるデータの相関ルールであるので，抽出されるデータは U2R 攻撃の兆候のデータである．

2 段階目では 1 段階目で抽出された兆候のデータから一定の閾値分後のデータを検知結果として出力する．出力されたデータのクラスラベルを参照し U2R 攻撃を検知することができたかを調査する．

5 実験結果と評価

5.1 4 つの方法による相関ルールの抽出とその評価

5.1.1 相関ルールの抽出

NSL-KDD データセット [3] の KDDTrain+ から適切なデータを取り出す．パラメータ m は U2R ラベルのついたデータから遡るパケット数を示す．本研究ではデータを取り出す方法として次の 2 つの方法を用いる．

ア U2R 攻撃ラベルのついたデータから m パケット前のデータのみを取り出す

イ U2R 攻撃ラベルのついたデータから前 10 パケットを取り出す

相関ルールを抽出する際データの属性を選択する．データの属性は次の 2 つの方法で選択し実験を行う．

1 U2R 攻撃の特徴が表れると考えられる属性を選択

2 訓練用データセットのほとんどのデータで同じ値である属性を除く属性すべてを選択

5.1.2 相関ルールを用いたテスト

KDDTest+ を h2 データベース[*1]に挿入し，相関ルールマイニングで抽出された相関ルールを用いて検索することで U2R 攻撃の兆候の検知を行う．抽出した相関ルールを使って検索すると U2R 攻撃の兆候とみられるデータが結果として得られる．

5.1.1 項で述べたデータセットからデータを取り出す際の 2 つの方法のうち，［ア］の方法を用いる際はテスト時では m パケット後のデータを出力する．［イ］の方法を用いる際はテスト時に k= {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} 代入し k パケット後のデータを出力する．ファイルに抽出されたデータのクラスラベルの名前が U2R 攻撃の種類の名称であれば，抽出された相関ルールは U2R 攻撃に有効なルールであるといえる．

5.2 NSL-KDD データセット

本研究の技術課題であるソフトウェアアーキテクチャの有用性の評価を NSL-KDD データセット [3] を利用して行う．本研究で NSL-KDD データセットを利用する理由は 2 点挙げられる．

1 点目に冗長性がないという特徴があることで学習性能に偏りが生じないことが挙げられる．重複レコードを持つと頻度の高いレコードに分類器の性能が偏ってしまうので，U2R 攻撃のような頻度の低いレコードの学習を妨げ，検知精度が損なわれてしまう．

2 点目に実際の通信に近い攻撃割合であることでデータを加工せずにフルセットで利用できることが挙げられる．正常通信に比べて攻撃通信の割合が多い場合，攻撃通信をカテゴリごとに抽出するなど工夫して使用する必要がある．フルセットで利用できることで評価結果には一貫性があるので等しいデータセットを使用する異なる研究と比較可能になる．

5.3 実験結果

本節での［ア］,［イ］,（1）,（2）はそれぞれ 5.1.1 項で述べた方法を表す．代表的な結果として実験［a］と［b］を記述する．実験［a］の実験結果を表 1，実験［b］の実験結果を表 2 に示す．

表 1 実験［a］における検知率

パケット数	再現率	偽陽性率	適合率	F 値	正解率
1	82.5%	78.7%	0.93%	0.0184	21.9%
20	84.0%	71.5%	1.04%	0.0205	29.0%
25	84.0%	85.2%	0.88%	0.0173	15.5%
30	80.5%	80.0%	0.89%	0.0176	20.5%

表 2 実験［b］における検知率

パケット数	再現率	偽陽性率	適合率	F 値	正解率
1	48.0%	44.6%	0.95%	0.0187	55.3%
20	41.0%	32.8%	1.11%	0.0215	67.0%
25	65.5%	65.5%	0.93%	0.0183	37.7%
30	21.5%	26.3%	0.73%	0.0141	73.3%

実験［a］では U2R 攻撃の兆候のデータ抽出方法として［ア］，属性の選択方法として（1）を用いた．m=20 の時は再現率がで 84% に達したが，偽陽性率は 71.5% であった．相関ルールを抽出する際，選択する属性が少なく相関ルールを見つけることができない場合があった．

実験［b］では U2R 攻撃の兆候のデータ抽出方法として［ア］，属性の選択方法として（2）を用いて行った．実験［a］より低い偽陽性率を達成することができたが，再現率も低下した．正解率は 60% 前後であり m=30 の時，73.3% と比較的高い数値を示している．これより U2R 攻撃は十分に検知できなかったが U2R 攻撃ではないデータを正しく検知できた割合が大きかったといえる．

4 つの実験結果を比較すると F 値において近い数値を示しており，2 つの値を除いてどれも 0.020 に満たない数値である．実験［a］と実験［b］の m=20 の時 F 値は比較的高い値を示したが再現率と偽陽性率が低いので効果的であるとはいえない結果である．

6 考察

実験結果より期待する検知結果を十分に得ることができなかった原因として次の 2 点が考えられる．

1 属性の選択に再考の余地があること

[*1] https://www.h2database.com/html/main.html

2 攻撃の兆候を調査する範囲に再考の余地があること

1つ目の原因は相関ルールマイニングを行う際の属性の選択に再考の余地があることである．リフト値が 1 以上である相関ルールをすべて抽出し，相関が強いといえる相関ルールを利用して検知を行ったが効果的に U2R 攻撃を検知することができなかった．ただし，それぞれの実験方法で比較的高い F 値を表している結果があるので，その実験で使用したルールを参考に再考することができると考える．

2 つ目の原因は U2R 攻撃の兆候データを調査する範囲に再考の余地があることである．コンピュータ侵入時の通信が一定の範囲内で行われていない場合，または 30 パケットより前にも兆候のデータが多くある場合，この実験では有効なルールを抽出することが困難である．なぜなら，今回の実験では U2R 攻撃から時系列的に一定の範囲を調査し，最大 30 パケット分前まで遡って調査を行っているからである．ただし実験［a］と実験［b］それぞれの適合率において m=20 の時比較的高い値を示しているので，20 パケット前に兆候のデータが表れやすく属性選択によって有効なルールを得られる可能性も考えられる．

アーキテクチャの有用性の評価について，テスト時に相関ルールマイニングのみを用いた場合と 2 段階アーキテクチャを用いた場合の検知数の結果を比較した．実験［a］と実験［b］の U2R 攻撃の検知数が高かった m={1，20，25} の場合と，反対に検知数が低かった m=30 の場合でも検証を行い，ほとんどの場合で相関ルールマイニングのみの場合よりも 2 段階アーキテクチャを使用して検知を行った場合の方が検知数が多くなった．この結果から 1 段階目では U2R 攻撃の兆候を得ることができており，2 段階目で U2R 攻撃を検知することができている．よって提案する 2 段階アーキテクチャは U2R 攻撃の検知に有用性があるといえる．

7　おわりに

ネットワーク攻撃に対して相関ルールマイニングを用いた侵入検知システムが盛んに使用されている．分析者はデータ間の属性値の相関性を見ることができるので，侵入検知システムの検知結果の理解性が高い．

相関ルールマイニングは侵入プロセスが複雑である攻撃に対する検知率が低いという課題がある．Brahmi ら [2] による研究では，U2R 攻撃は比較的低い検知率が報告されている．U2R 攻撃は様々な方法でコンピュータ内へ侵入し，最終的に管理者権限を奪う攻撃である．

本研究の目的は複雑なプロセスを持つ攻撃の侵入プロセスを考慮した侵入検知システムの設計である．本研究では複雑な侵入プロセスを持つ攻撃として U2R 攻撃に注目する．

技術課題としては「複雑な侵入プロセスを捉えることができるソフトウェア構造を明らかにすること」と「ソフトウェアアーキテクチャの有用性の評価」の 2 点がある．

それぞれの技術課題に対し，「プロセスの特徴に応じて役割分担を行う 2 段階アーキテクチャの提案」と「NSL-KDD データセットを用いた評価」を解決方法として提案する．

本研究では 2 段階アーキテクチャを提案する．トレーニング時は 1 段階目では侵入プロセスを考慮してデータを絞り込み，2 段階目で絞り込んだデータから相関ルールを抽出する．テスト時は 1 段階目にテスト用データセットから相関ルールに従ってデータを検出し，そのデータから一定の閾値分後のデータを検知結果として出力する．出力されたデータのクラスラベルを参照し評価を行う．

相関ルールに入力するデータの取り出し方法で 2 つ，属性選択の方法で 2 つを組み合わせた 4 つの実験を行う．

本研究では提案したアーキテクチャが有効であるといえる検知結果を十分に得ることができなかった．再現率が高い場合は偽陽性率も高くなり，偽陽性率が低いと再現率も低いという結果であり，U2R 攻撃の兆候の相関ルールを十分に抽出できたとはいえない．

実験結果から有効な相関ルールを得るためには，属性の選択と U2R 攻撃の兆候を抽出する範囲の 2 つを再考する必要があると考えた．

参考文献

[1] Anna L. Buczak, Member, IEEE, and Erhan Guven, Member, IEEE, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection”,IEEE Communications Surveys & Tutorials. Vol. 18, No. 2.

[2] H. Brahmi, B. Imen, and B. Sadok, “OMC-IDS: At the cross-roads of OLAP mining and intrusion detection”, Advances in Knowledge Discovery and Data Mining. pp. 13-24, 2012.

[3] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set”, Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[4] 高原尚志：“ネットワーク型侵入検知手法を用いた KDD Cup 1999 Data と Kyoto2016 の比較”， インターネットと運用技術シンポジウム論文集，pp. 77-84, 2018.

[5] 中田豊久：“基礎から学ぶデータマイニング”， コロナ社， 2013.

[6] 中村行宏．若尾靖和．林静香 ： “情報セキュリティの技術と対策がこれ 1 冊でしっかりわかる教科書”，技術評論社，2023.

[7] The University of Waikato, Weka, Version 3.9.6. “Weka 3: Machine Learning Software in Java”, https://www.cs.waikato.ac.nz/~ml/weka/index.html, 2023. (Accessed 2024.02.16)