3B3-OS-10a-4

外れ値検出に基づく対話的ファイアウォールログ分析

Interactive Firewall Log Analysis based on Outlier Detection

岡部正幸*1　山田誠二*2
Masayuki OKABE　Seiji YAMADA

*1豊橋技術科学大学
Toyohashi University of Technology

*2国立情報学研究所／総合研究大学院大学／東京工業大学
National Institute of Informatics

This paper considers a useful mutual feedback design for cooperative problem-solving between human and system on the task of outlier detection from network firewall log. We introduce a mutual feedback design to search partial feature spaces suitable for outlier detection efficiently.

1. はじめに

データ分析は，データに潜む構造・パターンに関する仮説を分析者が立案・設定し，機械学習を含む統計的手法による分析・検証を行い，その結果に対してパラメータ調整，特徴の追加・削除，仮説の再設定などのフィードバックを行うという一連の作業を繰り返しながら，人間とコンピュータが対話的に問題解決を図る作業といえる．この作業において価値ある分析結果を引き出すには，一般に分析者の能力によるところが大きいが，分析システム側からの支援機能，例えば高速計算，データ可視化などを提供することにより，仮説立案への手がかりを与えることができれば分析者にとって助けになる（図 1）．このように人間（分析者）とコンピュータ（分析システム）が協調して問題解決を行うには相互の能力を引き出すフィードバックを提供しあうインタラクション設計が不可欠である [岡部 13]．

本研究では，データ分析タスクの一例として，外れ値検出をベースとしたファイアウォールログ分析を取り上げ，タスク処理を効率的に行うために役立つ相互フィードバック設計について検討する．

2. ファイアウォールログ分析タスク

近年，セキュリティインシデントの発生は増加の一途をたどっており，コンピュータおよびネットワークセキュリティの重要性が高まっている．これらのインシデントを防ぐ，また発生した場合に迅速に対応するためには，ネットワークトラフィックログから異常通信を検知することが重要である．異常通信には，DoS 攻撃，ポートスキャン，ssh ブルートフォース攻撃など不正アクセスを行うため組織外から送信される場合のほか，組織内のホストがウイルス感染やクラッキングなどによって乗っ取られ，スパム送信などを大量に発生する場合などがある．また，P2P ソフトウェアによる通信なども著作権法違反になるデータの送受信を行っている場合が多く検知対象となる．

本研究では，組織内と組織外の境界に設置されたファイアウォールから出力されるネットワークトラフィックログをソースとして，異常通信の対象となっている組織内のホストを外れ値検出によって特定するタスクを考え，タスクを人とシステムが協調して効率的に解決するために役立つ相互フィードバック機能について検討する．

連絡先: 岡部正幸，豊橋技術科学大学情報メディア基盤センター
〒 441-8580 豊橋市天伯町雲雀ヶ丘 1-1
okabe@imc.tut.ac.jp

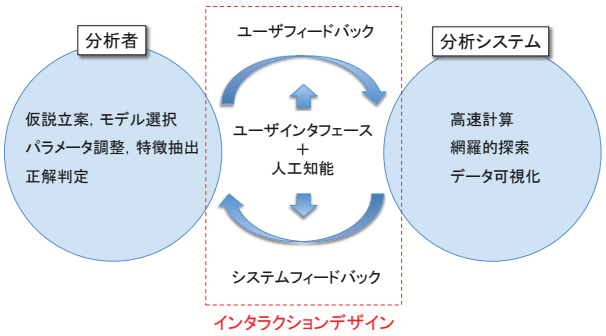


図 1: データ分析における相互フィードバック

2.1 ログデータ

本研究では，パケットがファイアウォールを通過した際に出力されるログデータを分析対象とする．図 2 に実際に出力されるログのフォーマットを示す．各ログには，パケットがファイアウォールを通過した日時，プロトコル，送信元ホストの IP アドレスとポート番号，送信先ホストの IP アドレスとポート番号が記されている．

2.2 特徴抽出

本研究で対象とするファイアウォールログ分析タスクの目的は，異常通信を送受信している組織内ホストの発見である．具体的には，DoS 攻撃，クラッキングを受けているホスト，ウイルス感染，ボットを仕込まれたホストなどを発見することである．このため，前節で説明したログから組織内ホストに関する特徴ベクトルを生成する．特徴ベクトルの属性として用いる特徴量は以下のものをベースとする．

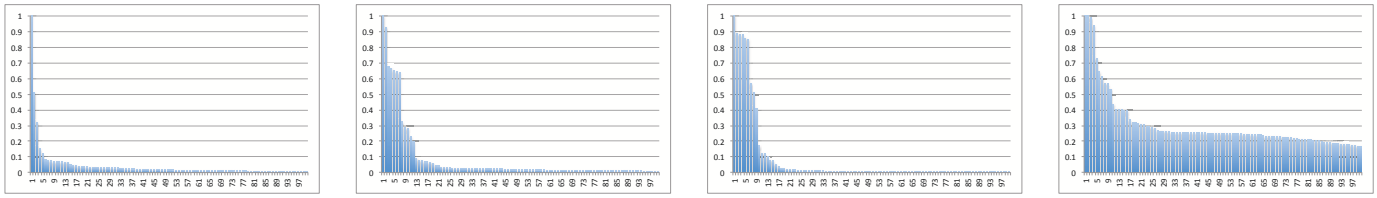
- 送信パケット数
- 送信先 IP アドレスの異なり数
- 送信元ポートの異なり数
- 送信先ポートの異なり数

上記の特徴量をベースに，更に細かな特徴量を生成することもできる．例えば，以下のような特徴を生成することができる．

1. 送信先 IP アドレスあたりの送信パケット数の平均と分散
2. 送信先 IP アドレスあたりの送信元ポート異なり数の平均と分散

```
Feb 03 2014 11:27:15: log_id: access-list acl_out permitted tcp inside/XX.XX.XX.XX(54732) -> inside/YY.YY.YY.YY(80)
```

図 2: ファイアウォールから取得したログデータ



(a) 部分特徴空間 1　(b) 部分特徴空間 2　(c) 部分特徴空間 3　(d) 部分特徴空間 4

図 3: 部分特徴空間における外れ値スコアをランキング順に並べた棒グラフ

3. 送信先 IP アドレスあたりの送信先ポート異なり数の平均と分散

送信元ポートおよび送信先ポートそれぞれについても同様な特徴量を生成できる．本研究では，このようにして生成した合計 22 次元の特徴ベクトルを各ホスト毎に生成する．

2.3 外れ値検出アルゴリズム

本研究では，k-近傍法に基づく外れ値検出方法をベースとする．この方法では，各データの k 番目に近いデータとの距離を外れ値スコアとして用い，このスコアが高いものから上位にランキングする [Chandola 09]．検出性能はパラメータ k によって変化するが，大きくは距離計算を行う特徴空間に依存する．前節では 22 種類の特徴量を利用するとしたが，それらがすべて役立つとは限らない．例えば，P2P 通信の検知に送信先 IP アドレスの異なり数は役立ちそうであるが，送信パケット数が役立つかどうかについては分からない．

近年，特徴空間を固定するのではなく，与えられた特徴量の中から外れ値検出に役立つ部分特徴空間を自動的に選別するアルゴリズムに関する研究が行われている [Müller 11]．特徴量の数を d とした場合にその組み合わせは $2^d - 1$ 個あるため，用いる特徴量の数が大きい場合には全探索は現実的ではない．このため，部分特徴空間に基づく外れ値検出では，外れ値の存在する特徴空間の探索（特徴量の組み合わせ探索）を効率的に行うことが求められる．

3. フィードバック設計

前節で述べたように k-近傍法に基づく外れ値検出では特徴空間の選択が検出性能に大きな影響を与えるが，特徴量の数が多いと検出に適した組み合わせを探索するのは容易ではない．通常，その探索は分析者の能力（専門知識やカン）に基いて行われるが，本研究では，分析者と分析システムが相互にフィードバックすることにより特徴量の組み合わせを効率的に探索することを試みる．具体的には，分析者が探索範囲を限定する情報をフィードバックする機能，また分析者が探索範囲を限定するための手がかりとして，各データの外れ値スコアをランキングリスト順に並べた棒グラフを部分空間毎に生成・提示する機能を追加する．図 3 に棒グラフの例を示す．グラフの横軸はランキングリストの順位，縦軸は外れ値スコアである．分析者はグラフにおける分布の一様性などを見て外れ値検出に役立ちそうな部分特徴空間を選択する．部分特徴空間の探索は以下のようにボトムアップに行う．

step1 与えられた特徴量の集合を V とする．また，$n = 1$ とする．

step2 V から生成可能な n 次元の特徴空間すべてにつき，各データの外れ値スコアをランキングリスト順に並べた棒グラフを生成する．

step2 分析者は役立ちそうな棒グラフを複数選択し，対応する外れ値ランキングリストそれぞれにおいて上位にランキングされたホストを調査する．

step4 選択された複数の棒グラフに対応する特徴空間を構成する特徴量（特徴軸）の和集合を計算し，これを V とする．$n = n + 1$ として，step2 に戻る．

step2 における分析者の選択により，特徴量の集合 V の要素数をうまく減らすことができれば，部分特徴空間の効率的な探索が行えると考えられる．

以上まとめると，本研究で提案するファイアウォールログからの外れ値検出タスクにおける分析者と分析システム間で行われる相互フィードバックは以下のようになる．

- 分析者側からのフィードバック：部分特徴空間の優先選択．
- 分析システム側からのフィードバック：各データの外れ値スコアをランキングリスト順に並べた棒グラフの部分空間毎の生成・提示．

4. まとめ

本研究では，ファイアウォールログ分析タスクにおいて，人とコンピュータが協調して問題解決を行うために役立つ相互フィードバック設計について検討した．分析手法として用いる k-近傍法に基づく外れ値検出において，外れ値スコアを計算するための部分特徴空間の探索を対話的に効率よく行うための提案を行った．今後，システム実装を進めるとともに提案手法の有効性について検証していく予定である．

参考文献

[Chandola 09] Chandola, V., Banerjee, A., and Kumar, V.: Anomaly Detection: A Survey, *ACM Computing Surveys*, Vol. 41, No. 3, pp. 15:1–15:58 (2009)

[Müller 11] Müller, E., Schiffer, M., and Seidl, T.: Statistical Selection of Relevant Subspace Projections for Outlier Ranking, in *Proc. of ICDE*, pp. 434–445 (2011)

[岡部 13] 岡部 正幸, 山田 誠二：知的インタラクティブシステムにおけるインタラクションデザインとは何か, 第 27 回人工知能学会全国大会, pp. 2F4–OS–04–5 (2013)