

確率的な推論方法を用いた異常検出セキュリティシステム

研川 幸雄[†] アントニ ローレンス[‡] ラシキア ジョージ[‡]

[†]岡山理科大学大学院 〒700-0005 岡山県岡山市理大町 1-1

[‡]岡山理科大学 〒700-0005 岡山県岡山市理大町 1-1

E-mail (toishigawa, anthony, lashkia)@ice.ous.ac.jp

あらまし 現在、セキュリティソフトの種類は多くある。しかし、アカウントの有効なユーザーや成りすましによる不正行為を検出するシステムはまだ少ない。本論文では異常検出に機械学習アプローチの1つである単純ベイズを使用し、新しいセキュリティシステムを提案する。このシステムは有効なユーザーの通常状態でのコマンド操作を学習することで現在のコマンド操作情報から異常状態を検出する。この作成したシステムによる不正侵入の検知、現在操作中であるユーザーの判別法の詳細及び評価について述べる。

キーワード セキュリティ、ネットワーク、データマイニング

A Security System for Anomaly Detection using Probabilistic Reasoning

Sachio TOISHIGAWA[†] Laurence ANTHONY[‡] George V. LASHKIA[‡]

[†] Graduate School of Okayama University of Science, Okayama University of Science, 1-1 Ridai-cho, Okayama
700-0005, Japan

[‡] Okayama University of Science, 1-1 Ridai-cho, Okayama 700-0005, Japan

E-mail (toishigawa, anthony, lashkia)@ice.ous.ac.jp

Abstract There are currently many kinds of security software. However, there are still few systems that can detect unauthorized actions by valid account holders or by camouflage. In this paper, we propose a new security system that can identify such actions by applying a novel variation of the Naïve Bayes machine learning algorithm to anomaly detection. In order to classify a user's behavior, the system uses a set of command operations of a target user to construct a profile of the users 'normal' operations. Subsequent profiles of the same or a different user's operations are then compared and classified accordingly. Results from applying the system to user distinction and anomaly detection demonstrate that it can be an effective and practical approach in real-world contexts.

Keyword Security, Network, Data Mining

1. はじめに

現在インターネットの普及により常時接続のブロードバンドの利用者が増えている。その中、ネットワークを構築している組織にとって、組織内のネットワークやホストが踏み台に利用される危険性が増している。また組織内のユーザーが他のユーザーのアカウントで不正行為をすることで組織内の安全性に問題が生じることは社会的信用性を失いかねない。そのため企業はファイアウォールやIDSなどのセキュリティシステム導入することで外部からの脅威を防いでいる。しかし、それらのセキュリティシステムは主に外部からのアクセ

ス等を監視、防御するものであり内部からの脅威に対してはあまりにも脆弱である。よって、アカウントの有効なユーザーに成りすました場合やそれにより踏み台にされた場合ネットワークの管理者がその不正を発見することは極めて困難である。このようなことにならないように、ユーザーのコマンド操作を監視することで現在操作しているユーザーが不正であるかを判別することができることが求められる。現在、このようなユーザーの操作を監視するシステムはあまり存在せず、機能も限られている。例えば、管理者があらかじめ指定したキーやプログラムを実行されたときに特定の動作を実行

させるだけとなる。本論文では、Linux 上でのシェルコマンド操作を学習データとしユーザーの操作を監視することで現在操作しているユーザーが正規のユーザーとの類似度を求めることで異常検出をする。この類似度を得るために Naïve Bayes[1][2]による機械学習アプローチを提案する。

2. 侵入検知の方法

2.1. 既存の侵入検知

今までにユーザーのコマンド操作を利用したセキュリティシステムは 1980 年代から研究が行われているが提案や推測などで当時あまり実用されていなかった。そして、1990 年代半ば頃から学習データを使用した実用的な研究が発表された。その中でもっとも実用的な研究が[3]である。この研究では、ユーザーごとのコマンド操作を記録しそのデータを学習データとし、学習データにあるコマンド順列と現在操作しているコマンド順列を比較し類似するたびに重みを与えることで学習データのユーザーと現在操作しているユーザーの類似度を求めている。そして、他のユーザーとの類似度を比較するためにグラフ化している。しかし、データにノイズがあるのでグラフが激しく振動する。このため、フィルタ処理をすることで滑らかにしている。つまり、この研究ではフィルタ処理を行った後の結果でないとユーザーの比較ができないという問題がある。

2.2. システム構成

我々が提案する異常検出セキュリティシステムは、機械学習の手法のひとつである Naïve Bayes[1][2]を用いる。Naïve Bayes の特徴により単純で計算が速く、ダイナミックに類似度を得ることができる。これにより学習データ (User Profile) との類似度をグラフにし比較することで不正行為の異常検出を行う (図.1)。

2.2.1. データの取得

異常検出をするための前準備として有効なアカウントのユーザーの通常状態でのコマンド操作を記録する。これをユーザーの学習データとする。Linux 上の Shell Window からコマンドキー操作を取得するのに Kernel Based Keylogger[4]を使用した。

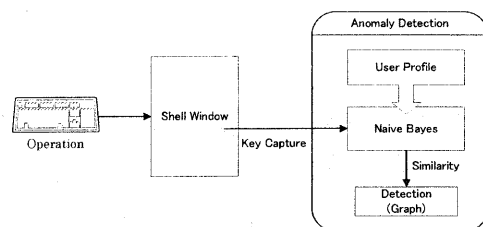


図.1 検知システム構成図

チャンク数 1	チャンク数 2
⋮	⋮
cd	cd /tmp
/tmp	/tmp ls
ls	ls perl
perl	perl -w
-w	-w test.pl
test.pl	test.pl clear
clear	⋮
⋮	

表.1

	有効数(個)	学習データ(個)	テストデータ(個)
User0	5588	3911	1677
User1	1509	1056	453
User2	2270	1589	681
User3	1551	1086	465

表.2

以下は取得したデータの一部である。

```

⋮
>cd /tmp
>ls
>perl -w test.pl
>clear
⋮

```

ここで、コマンド操作順序を以下のようにスペースで区切るようにする。

```

⋯ cd /tmp ls perl -w test.pl clear ⋯

```

これをコマンド順列とする。また、現在操作しているユーザーのコマンド操作に対しても同じようなコマンド順列を作る。記録した有効なアカウントのユーザー操作の学習データとそれ以後の操作しているユーザーのコマンド順列の類似度

を求める。また、グラフにするためにはユーザーのコマンド順列を任意の長さに取り、固定した長さのシーケンスに分ける必要がある。

2.2.2. 学習データとの類似度

類似度を得るための Naïve Bayes[1][2]の使用について説明する。Naïve Bayes アプローチによる式は

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j) \quad (1)$$

である。ここで、 $V = \{v_1, v_2, \dots\}$ は有限集合である学習データである。 $P(v_j)$ は学習データの確率分布。および $\prod_i P(a_i | v_j) = P(a_1, a_2, \dots, a_n | v_j)$ はコマンド順列の確率分布である。また、 i はコマンド順列に存在する語数である。しかし、今回の実験では1ユーザーに対して1学習データしかないため $P(v_j)$ の学習データの確率分布は1となる。式(1)を用いたシステムにより現在操作しているユーザーのシーケンスとの類似度の変化をグラフ化し比較することで不正行為の異常検出を行う。また、このシステムでは以下のパラメータを使用する。

- ・ シーケンス語数 (i): 学習データと操作しているユーザーの類似度を得るためのシーケンスの長さ。この値を大きくすることで利用するユーザー特徴が増える。しかし、値を大きくしすぎると類似度を得るのに時間がかかってしまう問題がある。
- ・ ステップ数: ユーザー操作の時間変化を調節するためにシーケンスをずらす長さ。この値を小さくすると前後の類似度の変化が小さくなり、また大きくすると急激に変化する場合がある。
- ・ チャンク数: 連続したコマンドを1組のキー操作とする長さ。例えば、表.1で示しているように、チャンク数が増えることで、よく使うコマンドの組み合わせ特徴が得られる。

3. 評価実験

我々の異常検出システムを評価するために、研究室にいるユーザー4人に約3ヶ月間のLinuxでのシェルコマンドの履歴を取得した。取得した結果、それぞれの有効コマンド順列の長さはUser0が5588個、User1は1509個、User2は2270個、User3は1551個である(表.2)。また、これらのデータは先に述べたように各ユーザーの普段行っている操作を記録

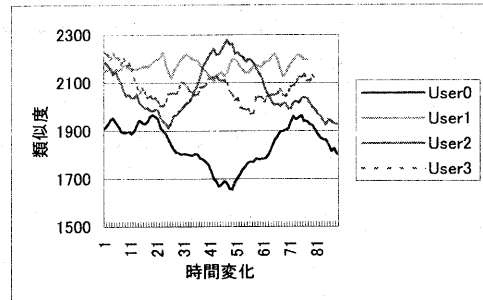


図.2 チャンク数1 (学習データは User0 を使用)

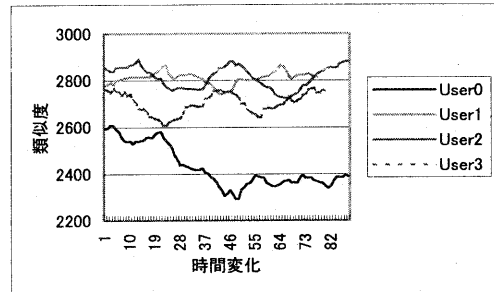


図.3 チャンク数2 (学習データは User0 を使用)

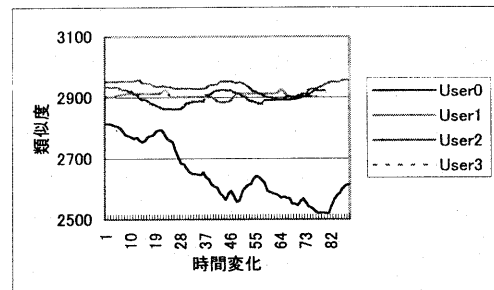


図.4 チャンク数3 (学習データは User0 を使用)

したものである。そして、実験を行うためには学習データとテストデータが必要となる。このため、取得しているデータを機械的に最初の語から70%までを学習データとし、残りの30%をテストデータとして区切った(表.2)。例えば、User0のコマンド順列の長さは学習データが3911個、テストデータが1677個である。今回はUser0の学習データを使用して異常検出システムの結果をグラフにして評価を行う。各パラメータの値を変化させ最適な値を検討した(シーケンス語数(i)は100から500の間、ステップ数は5から30の間、チャンク数は1、2、3)。そして、実験により、シーケンス語

数 (i) = 350、ステップ数=15 となった。図.2、3、4 は User0 の学習データでそれぞれチャンク数 1、2、3 を使ったときの結果を示している。グラフは縦軸が類似度であり数値が小さくなるほど似ていることとなる。また、横軸がコマンド操作経過である。チャンク数 1 (図.2)、チャンク数 2 (図.3) と比べチャンク数 3 (図.4) の結果では、他のユーザーとの違いが大きくなっていることがわかる。また、各ユーザーの学習データのグラフも同様にチャンク数 3 の方がチャンク数 1 とチャンク数 2 のときよりも違いがより明確になっている。しかし、チャンク数 3 のみを使うとユーザー特徴が場合によって極端に減ってしまうこともある。そこで、チャンク数 1、2 との組み合わせにより、必要なユーザー特徴を確保でき、そのうえ他のユーザーとのグラフ線との違いがさらにはっきりと現れている (図.5、図.6)。

4. まとめ

評価実験の結果、今回作成した異常検出セキュリティシステムによって学習データと違うユーザーが操作することで正規のユーザーとの違いを検知することを確認した。また、同一のユーザーであっても操作に変化があることを確認した。また、グラフで表しているように始まりの時点ですでに他のユーザーとは大きな違いがあるので不正行為が初期の段階であったとしても検知できることを表している。すなわち、ネットワーク等のシステム導入初期のセキュリティホールが多くある状態での不正行為を検知できる。これはより安全にシステムを構築できるということに有効である。今回の実験では、1つのホストだけであったが1つ1つのホストのセキュリティを確保できることはネットワーク全体のセキュリティにつながることをいえる。今後の課題としては、コマンド操作だけでなくシステムコールによる不正検出機能を加える。特定の操作に対しての重みを与えることでより不正行為の検出の精度を高める。さらにサーバー等による統括管理ができる拡張性・柔軟性をもつセキュリティシステムへの考察がある。

文 献

- [1] Tom M. Mitchell "Machine Learning", pp.177-184.
- [2] 矢入, "Machine Learning 輪読会",

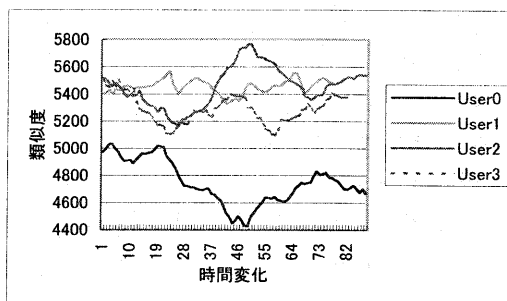


図.5 チャンク数 1、2 の組み合わせ
(学習データは User0 を使用)

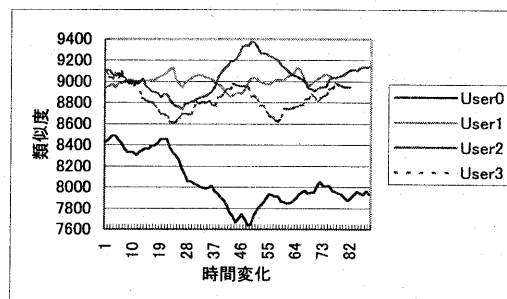


図.6 チャンク数 1、2、3 の組み合わせ
(学習データは User0 を使用)

http://ken.ai.rcast.u-tokyo.ac.jp/~yoshi/ML_Seminar/ML_Chap6.pdf, 2000.

- [3] Terran Lane and Carla E. Brodley, "An Application of Machine Learning to Anomaly Detection" February 14, 1997.
<http://citeseer.nj.nec.com/lane97application.html>
- [4] Mercenary, "Kernel Based Keylogger",
<http://packetstormsecurity.nl/UNIX/security/kernel.keylogger.txt>, 2002.
- [5] 平岡 真崎、石田 賢治、天野 橘太郎, "インターネット不正アクセス「踏み台」検知システム", 第3回 IEEE 広島支部 学生シンポジウム論文集 pp.143-146, 2001.
- [6] 浅香 緑、女部田 武史、井上 直、岡澤 俊士、後藤 滋樹, "不正侵入の痕跡と判別分析によるリモートアタックの検出法", 電子情報通信学会論文誌, Vol.J85-B No.1, pp.60-74, 2002.
- [7] 武田 圭史、磯崎 宏, "ネットワーク侵入検知" pp.109-125, 2000.