単純ベイズを用いたユーザー操作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　‡岡山理科大学

1. はじめに

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ある。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を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、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。本論文では、Linux 上でのシェルコマンド操作を取得しユーザーの操作を監視する。そして、現在操作しているユーザーと学習データの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統計的な機械学習アプローチを提案する。

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。まず、異常検出をするために必要な学習データ（User Profile）を取得する。取得方法は Shell Window からのコマンドキー操作を Kernel Based Keylogger[3]で記録する。記録したデータのコマンド操作順序を以下のようにスペースで区切る。

… 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。このとき、ユーザーのコマンド順列は任意の長さを取り、固定した長さのシーケンスに分ける必要がある。類似度は Naïve Bayes の特徴により単純で計算が速く、ダイナミックに得ることができる。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。このシステムでは以下のパラメータを使用する。

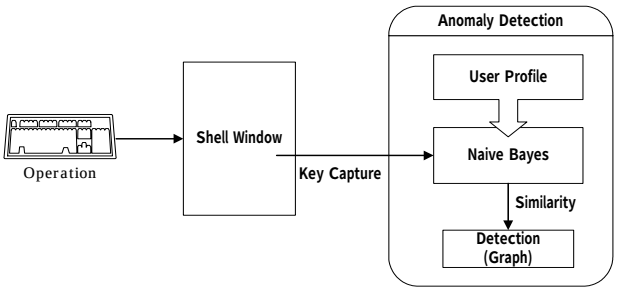

図.1　検知システム構成図

	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.1

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.2

- シーケンス語数：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複数個で１つの操作とする長さ。例えば、表.2 のようになる。チャンク数が増えることで、よく使うコマンドの組み合

Title: A Security System for User Operation Anomaly Detection using a Naïve Bayes Approach

†Sachio Toishigawa, Graduate School of Okayama University of Science

‡Laurence ANTHONY , Okayama University of Science

‡George V. LASHKIA, Okayama University of Science

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操作履歴を取得した。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。各データの個数は表.1 で示す。図.2、図.3 のグラフは User0 の学習データを使用して異常検出システムの結果を表示している。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）。そして、実験により、シーケンス語数=350、ステップ数=15 となった。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の時変化である。User0 の学習データのチャンク数 1（図.2）とチャンク数 2（図.3）では、チャンク数 2 の方が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 2 の方がチャンク数 1 のときよりも違いがより明確になっている。

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の操作が正規のユーザーとは異なることを確認した。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学習データのユーザーと異なるユーザーとは大きな違いがある。これは、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つながる。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といえる。今後の課題としては、チャンク数を増やす、またチャンク数の組み合わせによりユーザー特徴の変化を調べる。そして、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。

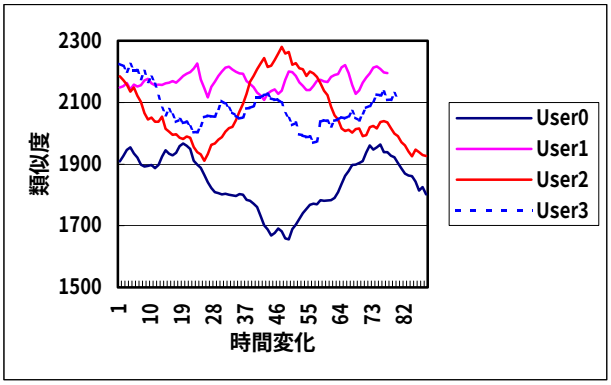

図.2　チャンク数 1

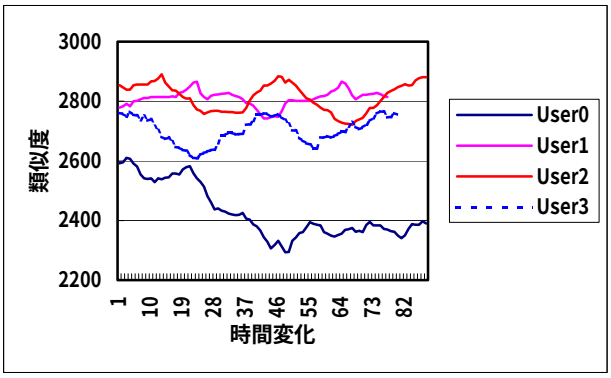

図.3　チャンク数 2

図.2，図.3 の学習データに User0 を使用している。

文　献

[1] Tom M. Mitchell “Machine Learning”, pp.177-184.

[2] 矢入, ”Machine Learning輪読会”, http://ken.ai.rcast.u-tokyo.ac.jp/~yoshi/ML_Seminar/ML_Chap6.pdf, 2000.

[3] Mercenary, ”Kernel Based Keylogger”, http://packetstormsecurity.nl/UNIX/security/kernel.keylogger.txt, 2002.

[4] Terran Lane and Carla E. Brodley, ”An Application of Machine Learning to Anomaly Detection” http://citeseer.nj.nec.com/lane97application.html, February 14, 1997.

[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.