2004－DBS－134　(I)　(12)
2004／7／13

時系列データからの多層ネットワーク特徴抽出手法の提案：Eigen Co-occurrence Matrix (ECM)

岡 瑞起† 　小磯 知之† 　加藤 和彦‡

†筑波大学大学院理工学研究科〒305-8573　茨城県つくば市天王台 1-1-1
‡筑波大学システム情報工学研究科 〒305-8573　茨城県つくば市天王台 1-1-1
E-mail:　{mizuki, koiso, kato}@osss.is.tsukuba.ac.jp

あらまし　コンピュータのセキュリティを考える上で重要な問題の 1 つに、有効なユーザに成りすますことによる不正行為を検知することが挙げられる．不正行為を検知する方法としては、異常検知によるアプローチが有効である．異常検知は、有効なユーザの挙動を学習することによりユーザのモデルを作成し、そのモデルから逸脱する挙動を異常と検知する．本稿では、異常検知に用いられる時系列データからのユーザの挙動の特徴抽出に着目し、Eigen Co-occurrence Matrix (ECM)手法という新たな時系列データからの特徴抽出手法を提案する．ユーザの UNIX コマンド時系列から ECM 手法を用いて特徴抽出を行い、異常検知に利用する．Schonlau らが提供する UNIX コマンドデータに対して成りすまし検知の実験を行った．

キーワード　成りすまし検知, 時系列データ, 共起行列, 主成分分析

Eigen Co-occurrence Matrix (ECM): Method for Extracting Features of Sequential Data as Layered Networks

Mizuki OKA† 　Tomoyuki KOISO† 　and 　Kazuhiko KATO‡

† Master's Program in Science and Engineering at University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan
‡ Graduate School of Systems and Information Engineering　1-1-1 Tennodai, Tsukuba, Ibaraki, 305-8573 Japan
E-mail:　{mizuki, koiso, kato} @osss.is.tsukuba.ac.jp

Abstract. One of the problems of importance in computer security is to detect the presence of an intruder masquerading as the valid user. Anomaly detection is a promising approach to detect intruders (masqueraders). Anomaly detection creates a user profile and labels any behavior that deviates from the profile as anomalous. A challenging task in detecting intruders is to model a user's behavior based on sequential data, which can be used to effectively distinguish anomalous behaviors. In this paper, we propose a novel method, called the Eigen Cooccurrence Matrix (ECM), that models sequences of user actions (UNIX commands) and extracts their principal features. We applied ECM on the experiment of masquerade detection framed by Schonlau et al. We report the obtained result from the experiment and compare it with those from several conventional methods.

Keyword　Masquerade Detection, Sequential Data, Co-occurrence Matrix, Principal Component Analysis

1. はじめに

不正アクセスによるファイルの改竄、情報漏洩、踏み台、成りすましによる被害が後を絶たない．不正アクセスを早期に発見し対策を講じるには、侵入検知システムの利用が有効である．侵入検知システムは検査対象の挙動を監視し、異常が発生したと判断されるとアラームを発生させたり、システムを停止させたりするなどの処理を行う．

本研究の目的は、UNIX コマンドを監視することによりユーザの動的な挙動をモデル化し、成りすましによる攻撃を防ぐ侵入検知システムの構築である．本研究では侵入検知システムの中の特に異常検知システムと呼ばれるシステムを対象とする．異常検知システムは正規のユーザの挙動を表すモデルを作り、現在操作

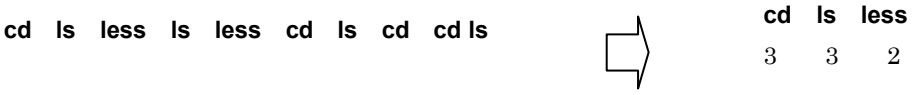

図 1. Histogram

図 2. Ngram (N=3)

しているユーザがそのモデルに従っているかどうかを検査することにより成りすましを検知する．異常検知システムは未知の異常な挙動を検知できるという特徴を持つ．

成りすましを検知する異常検知システムの構築の際、UNIX コマンドのようなユーザの挙動を表す時系列データがモデル化に良く用いられる．時系列データから特徴抽出を行い、有効なユーザか成りすましの識別を行う．典型的な特徴抽出の方法は、データに現れるイベントの Histogram や Ngram により特徴ベクトルに変換する[1,2,3]．しかしこれらの方法には、時系列データにおけるユーザの挙動の動的情報が失われるという問題や、隣接するイベント間の特徴しか表現されないという問題がある．

これら問題に対処するために、本稿ではユーザの挙動の動的情報をとらえ時系列データの特徴を抽出する Eigen Co-occurrence Matrix (ECM)手法を提案する[7]．ECM では、時系列情報を考慮しながら、イベント間の関連付けを行う．この関連付けは、二項間イベントに着目し全ての二項間イベントの関連性を Co-occurrence Matrix (共起行列)として表現することにより行う．共起行列は、全ての二項間の関連性の強さがその距離と出現頻度により表現されることになる．

正当なユーザと成りすまし者を識別するには、さらに共起行列をパターンとして扱い、統計的パターン認識手法を適用することが妥当と考えられる．最も簡単なパターン認識手法は、パターン間のマッチングに基づく手法であるが、共起行列そのものをパターンとして扱った場合、パターンの次元が膨大になってしまう．そのため、パターン間のマッチングでは、特徴を抽出し（情報圧縮にもなっている）、認識を行うことがより有効である．パターンから有効な特徴抽出を行うことにより、入力パターンの変動に対して頑健な認識結果が期待できる．

我々の提案する ECM 手法はこの特徴抽出手法として、主成分分析を用い、共起行列の識別に利用する．主成分分析はベクトル形式のデータを少数の特徴（主成分）で表すことを可能とする統計的手法である．主成分分析を用いた認識の成功例として、Turk ら[4]が提案した Eigenface (固有顔)による顔画像の認識が広く知られている．ECM は、Co-occurrence Matrix (共起行列)を顔画像と見なしたところに方式考案の着眼点がある．ECM 手法は主成分分析を介することにより、固有顔に対応する固有共起行列(Eigen Co-occurrence Matrix)を作成し、もとの共起行列を低次元で近似して表現することが可能である．ECM 手法はさらに、この近似された共起行列の二項関係をつなぎ合わせ自動的にネットワークを構築する．時系列データから特徴的なネットワーク構造を自動的に抽出できるという点は従来手法にない ECM 手法の特徴である．

本論文の構成を以下に示す．2 章で既存の特徴表現手法を説明する．3 章で我々の提案する ECM 手法について述べる．4 章で実験結果を示す．5 章でまとめと今後の課題を述べる．

2. 既存の特徴表現手法

時系列データを Histogram や Ngram により特徴ベクトルとして特徴を表現する手法が多く用いられている[1,2,3]．これらの手法を用い、ベクトルとして特徴を表現する利点は、ベクトルに対し主成分分析、ベイズ識別関数といった様々な数学的手法が適用可能なことである．

これらの手法を説明するための例として、ファイル参照を行っているユーザの挙動、**cd ls less ls less cd ls cd cd ls**、を考える．この時系列に対して Histogram により特徴表現を行うと図 1 のようになる．Histogram は、時系列に現れる固有のイベント列(例では **cd ls less**)の頻度を数えてベクトルとする．しかし、Histogram による特徴表現では、時系列情報を一切失ってしまうという難点がある．一方、Ngram (N=3) による特徴表現は図 2 のようになる．Ngram は、隣接する時系列イベントのみを特徴として表現するため隣接していないイベントの関連性は考慮することができない．例の時系列データでは、例えば、**ls less ls** というイベント列の関連性は捕らえられるが、**ls less less** とい

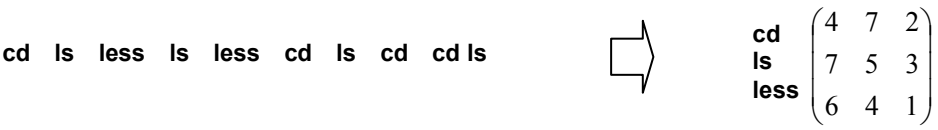

図 3. 共起行列

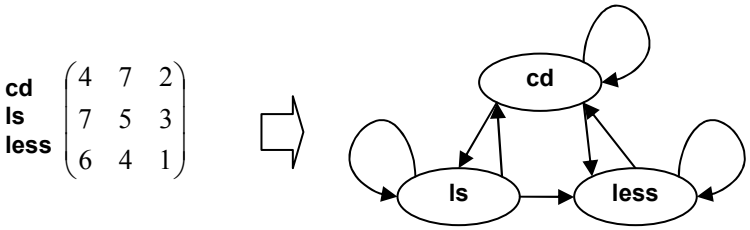

図 4. 共起行列のネットワーク表現

うイベント列の関連性は Ngram として現れないため捉えることができない．これば人間の流動的な行動を考えると制約の厳しすぎる特徴表現であると我々は考える．

3. 提案手法

3.1. Co-occurrence Matrix

我々の提案する ECM 手法は、時系列の特徴表現を Co-occurrence Matrix (共起行列)として行うことにより前章に述べた難点を解決する．共起行列とは、出現するイベントのそれぞれの二項間の関連性の強さを、ある距離の間に現れるイベントの出現頻度により表し、全ての二項間のイベントの関連性を表現した行列である．つまり、各二項間イベントの関連性の強さは、二項間の距離と出現頻度により表されることになる．

図 3 に 2 章と同じ時系列データ例を、共起行列として特徴を表現した結果を示す（関連性を考慮する距離を 6 つ先のイベントまで)．イベント列 **ls less ls** 関連性の強さは、**ls less** と **less ls** それぞれの関係性の強さである 3 と 4 で表現される．また、イベント列 **ls less less** の関連性の強さは、**ls less** と **less less** それぞれの関連性の強さである 3 と 1 で表現されることになり、Histogram, Ngram では全く捉えることができなかった特徴表現が可能になる．このように、時系列を共起行列で表現することにより人間の流動的な行動のモデル化が可能になる．

さらに、共起行列を隣接行列として捉えると、共起行列からのネットワーク構造が抽出される．図 4 に図 3 の共起行列をネットワーク構造として表現した結果を示す．

3.2. 主成分分析

時系列データに対して特徴表現を行った共起行列を識別するには、さらに共起行列をパターンとして扱い、統計的パターン認識手法を適用することが考えられる．しかし、共起行列そのものをパターンとして扱った場合、パターンの次元が膨大になってしまうため特徴抽出を行うことが必要になる．ECM 手法は、共起行列から特徴抽出を行うために主成分分析を用いる．

主成分分析とは多変量で表されるデータの統計から、一次結合で表現される新たな変量を構成し、互いに無相関な「主成分」に要約する手法である．主成分分析を用いた認識の成功例として Turk ら[4]が提案した Eigenface（固有顔)による顔画像の認識が広く知られている．ECM は、Co-occurrence Matrix（共起行列）を顔画像と見なし、Eigenface に対応する Eigen Co-occurrence Matrix（固有共起行列)を作成する．

共起行列からの主成分分析を用いた特徴ベクトル抽出は次に述べる手順で行う．

- p 枚の学習用の共起行列のうち i 番目の共起行列を、各要素の値を並べた N 次元のベクトル x_i として表現する．
- p 枚の共起行列の平均ベクトルを

$$\bar{x} = \frac{1}{p}\sum_{i=1}^{p} x_i \qquad (1)$$

とし、各共起行列から平均ベクトルを引いたベクトルを $\tilde{x}_i = x_i - \bar{x}$ で表し、各共起行列から平均ベクトルを引いた共起行列の集合を行列 $\tilde{X} = [\tilde{x}_1, \cdots, \tilde{x}_p]$ で表す．

- 学習用の共起行列集合を最適に近似する正規直行基底 a を、$\tilde{X}$ の共分散行列の固有ベクトルで構成する．このとき、a の各固有ベクトル a_l を、固有共起行列(Eigen co-occurrence matrix)とする．
- ある共起行列 x に対する主成分スコア C を x と a の内積を計算することにより求める．C の各成分 $c_1, c_2, \cdots, c_N$ は、共起行列 x を表現するための各固有共起行列の貢献度を表す

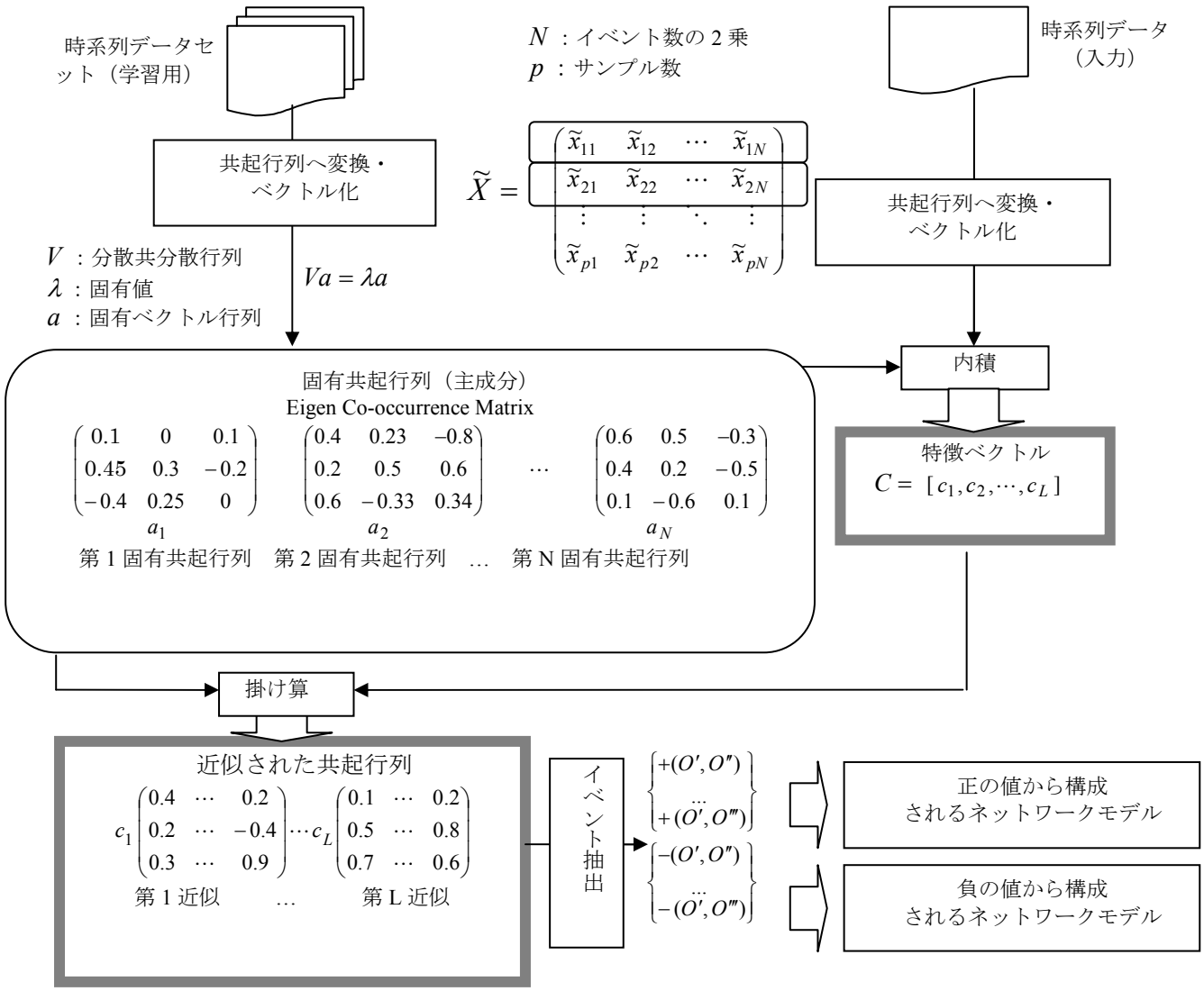

図 5. ECM 手法の全体図

ことになる．C ベクトルを共起行列の特徴ベクトル x とする．

3.3. 多層ネットワーク表現

固有ベクトルの次元 L（$L = 1, \cdots, N$）を小さくすることにより、固有共起行列 a と特徴ベクトル C を用いて、もとの共起行列を

$$\hat{\tilde{x}} = \sum_{i=1}^{L} c_i a_i \quad \text{for} \quad (L = 1, \cdots, N) \qquad (2)$$

のように低次元で近似して表現することができる．

また、固有ベクトルの次元近 L を変化させることにより、L 番目の近似要素におけるそれぞれの近似共起行列が

$$z_i = c_i a_i \quad \text{for} \quad (i = 1, \cdots, N) \qquad (3)$$

から生成できる．各層の近似共起行列からネットワークを抽出することにより、多層ネットワーク表現が可能である．それぞれの層のネットワークは、もとの共起行列の部分ネットワークではなく、固有共起行列から発生する全体の構造をもつ近似ネットワークである．

さらに、構成要素 z_i は、

$$z_i = c_i a_i = x(i) + y(i) \qquad \text{for} \quad (L = 1, \cdots, N) \qquad (4)$$

のように、正($x(i)$)と負($y(i)$)の要素を分離してそれぞれからは 1 つ 1 つのネットワークが構成できる．正の要素から成る行列 $x(i)$ のつくるネットワークは共起性が正の値でもって、(入力－平均) の行列を再構成するのに寄与し、$y(i)$ は同じく負の値でもって再構成に寄与するという違いがある．

以上の ECM 手法における一連の流れを図 5 に示す．

4. 実験

提案した ECM 手法に基づき異常検知システムを実装し、実際の UNIX コマンド時系列のログデータにおいて正常なユーザと成りすまし者を識別する実験を行った．

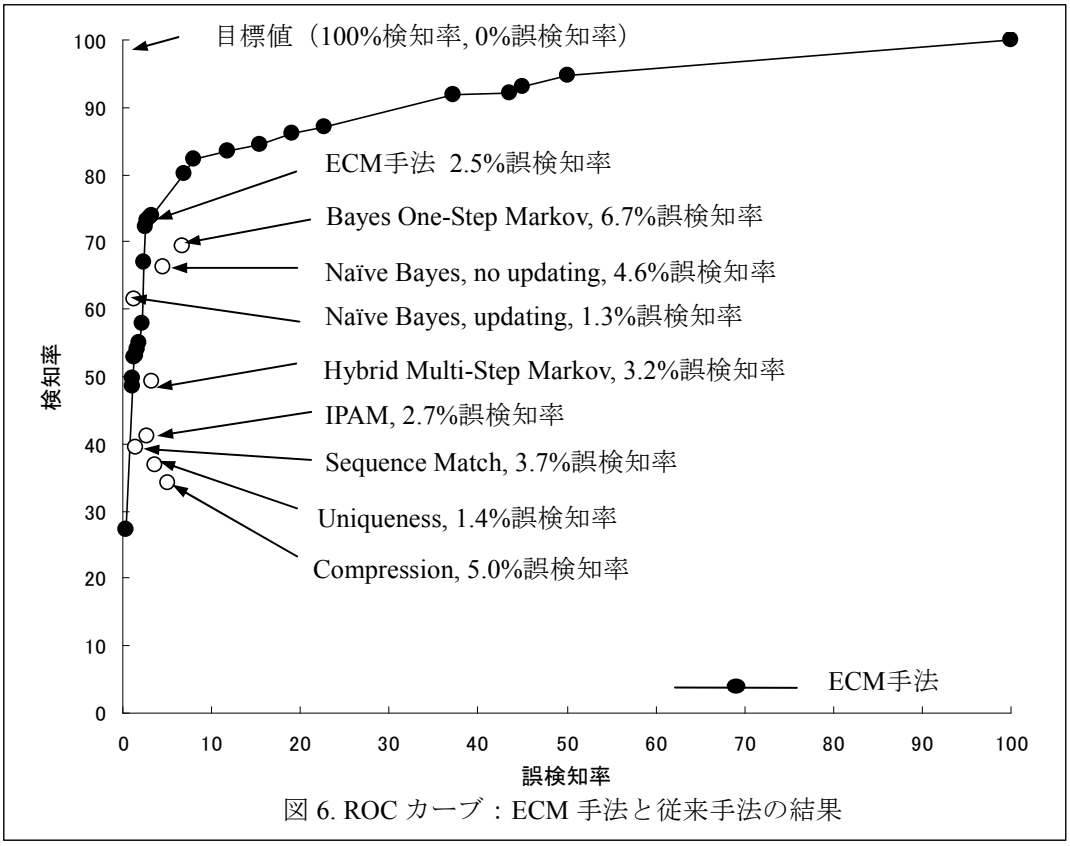

図 6. ROC カーブ：ECM 手法と従来手法の結果

4.1. データ

実験にはSchonlauら[5]が提供しているUNIXコマンドのデータを用いた．Schonlauらのデータには、1人のユーザにつき、15,000のUNIXコマンドの履歴が提供されており、それが50人分用意されている．彼らのデータには、プライバシーの理由から引数、フラグやエイリアスの情報は含まれていない．15,000コマンドのうち最初の5000コマンドは、正規のユーザのコマンドで構成されており、残りの10,000コマンドに成りすましのデータが挿入されている．最初の5000コマンドを学習データ、残りの10,000コマンドをテストデータとした．

4.2. 特徴抽出

検査対象は 100 コマンドずつとし、5,000 コマンドの学習データを 100 コマンドごとのウィンドウに分け、それぞれから ECM 手法を用いて特徴抽出を行う．

50 人全てのユーザの学習データを用い（50×50=2500 ウィンドウ）,固有共起行列を作成した．そのうち降順に並べられた最初の 50 個の固有値に対応する 50 個の固有ベクトルを固有共起行列として用いた ($N = 50$).

学習データ：各ユーザの学習データから 100 コマンドを 1 ウィンドウの単位とし、ECM 手法を用いてネットワークモデルに変換した．

テストデータ：各ユーザのテストデータから同様に 100 コマンドを 1 ウィンドウ単位とし、ECM 手法を用いてネットワークモデルに変換した．

4.3. 類似度

学習データネットワークモデルセットを S とし、それとテストデータ seq_i のネットワークモデルの類似度を計算し、「正常」か「異常」であるかを判断する．seq_i の学習データとの類似度は、式(5)で表わされるように、学習データそれぞれのネットワークモデルとの類似度で 1 番大きい値とする．

$$Sim(seq_i, S) = \max seq_{j \in S} \{Sim(seq_i, seq_j)\} \quad (5)$$

ここで、ネットワークモデルの類似度 $Sim(seq_i, seq_j)$ は、

$$Sim(seq_i, seq_j) = \sum_{k}^{N} \delta(T_k(i), T_k(j)) \quad (6)$$

のように定義した． $T_k(x)$ は seq_x が k 層においてつくるネットワークモデルを示し、$\delta(T_k(i), T_k(j))$ は、ネットワークモデル $T_k(i)$ と $T_k(j)$ が同一階層同士における部分ネットワークの一致数の和を示している．各層におけるネットワークは、対応する近似共起行列から値の大きい順に 30 個ノードを取り出し、構成した．

また、ノードが 3 つ繋がっているネットワークを 1 つの部分ネットワークと捉えた.

4.4. しきい値

ユーザ i ごとに、テストデータ seq_i が「正常」であるか「異常」であるか判断する類似度のしきい値 ε_i を設け、$Sim(seq_i, S)$ がしきい値 ε_i よりも大きければ正常、小さければ異常と判断する．しきい値 ε_i を変化させることにより検知率（異常な実行を異常と判断する）と誤検知率（正常な実行を異常と判断する）が変化する.

4.5. 実験結果

実験の評価には Receiver Operating Characteristic (ROC) カーブを用いた．ROC カーブとは縦軸に検知率、横軸に誤検知率をとり、しきい値を変化させたときの結果をプロットしたシステムの精度を表すグラフである．プロット点が図の左上に近ければ近いほど、誤検知率が低く、検知率が高いことを示し、性能が良いことを表す.

Schonlau ら[5]と Maxion ら[6]は本研究で使用した同様のデータセットに対し Bayes 1-Step Markov, Hybrid Multi-Step Markov, IPAM, Uniqueness, Sequence-Match, Compression, and Naïve Bayes と呼ばれる手法を適用している．ECM 手法を用いてユーザ i ごとに、しきい値 ε_i を変化させ 50 人分の結果をまとめた結果を彼らの結果とともに、図 6 に示す．図 6 の結果が示すように、我々の提案する ECM 手法が最も高い検知率の中で、最も低い誤検知率を示しており、手法の有効性が確認できた.

5. まとめと今後の課題

時系列データの特徴抽出を行う新たなECM手法の提案し、人間の挙動を解析することによる侵入検知システムに適用した．実用的な侵入検知システムは、検知率が 99%以上、誤検知率が 1%以下であることが不可欠であると言われている．我々の提案する ECM 手法を用いることにより既存の手法よりも良い精度を得ることができたが、目標とする精度にはまだ及んでいない.

今後は、共起行列を作成する際に、注目する二項間イベントが現れる距離に重みをつけ、時系列の特徴を表現し実験を行い、精度を比較したい．また、特徴表現を行う際、UNIX コマンドの引数、エイリアスなどの情報を利用していないが、今後それらの情報を活用して検知精度を高めて行きたい.

文　献

[1] N. Ye, X. Li, Q. Chen, S. M. Emran, M. Xu, Probablistic Techniques for Intrusion Detection Based on Computer Audit Data, IEEE Transactions of Systems Man and Cybernetics, Vol.31, pp.266-274, 2001

[2] S. A. Hofmeyr, S. Forrest, A. Somayaji, Intrusion Detection using Sequences of System Calls, Journal of Computer Security, vol.6, pp.151-180, 1998

[3] W. Lee, S. J. Stolfo, A framework for constructing features and models for intrusion detection systems, Information and System Security, vol.3, pp.227-261, 2000.

[4] M. Turk, A. Pentland, Eigenfaces for Recognition, Journal of Cognitive Neuroscience, vol.3, no.1, 1991.

[5] M. Schonlau, W. Dumonchel, W. H. Ju, A. F. Karr, M. Theus, Y. Vardi, Computer intrusion: Detecting masquerades, Statistica Science, vol.16, no.1, pp.58-74, 2001.

[6] R. A. Maxion, T. N. Townsend, Masquerade Detection Using Truncated Command Lines, Proc. International Conference on Dependable Systems and Networks (DSN-02), pp.219-228, Washington, 2002.

[7] M. Oka, Y. Oyama, H. Abe, K. Kato, Anomaly Detection Using Layered Networks Based on Eigen Co-occurrence Matrix, Proc. of 7th Int. Symp. On Recent Advanced Intrusion Detection (RAID), Sophia Antipolis, French Riviera, France, Sep. 15-17, 2004. (Accepted for publication).