2K-01 フリークエントエピソード解析を用いたユーザ監査ログデータからの異常侵入検出*

伊藤大介† 大和田勇人†

東京理科大学 理工学部‡

1 はじめに

異常侵入検出法は、過去の監査履歴と現在における振舞いの違いを比較することにより検出を行うが、[1] はその比較の基準としてイベントシーケンスの" 類似性" を用いている。時間軸をもった離散的なイベントの集合であるユーザ監査ログデータからクラスタリングを行い、各ユーザの分類に有効であることを示したが、同時にこの類似性の観点だけでは精度が不十分であるとしている。

そこで我々は、より正確な分類を行うためにシーケンス上におけるイベントの "頻出性" に着目した。類似性の観点だけでは各ユーザの特徴付けが不十分であり、振舞いの違いを認識するに至らない。そこでフリークエントエピソード解析 [2] により得られたルールを各ユーザの振舞いの違いとして考慮し、そこから類似度を求めることにより、分類の精度向上を図った。

2 方法

2.1 類似度の算出

[1] による類似度の算出方法は以下のようである。シーケンスの長さを l、シーケンス $X=(x_0, x_1, ...x_{l-1})$ と $Y=(y_0, y_1, ..., y_{l-1})$ に関する類似尺度の定式化は以下のようになる。

$$w(X, Y, i) = \begin{cases} 0 & \text{if i} < 0 \text{ or } x_i \neq y_i \\ 1 + w(X, Y, i-1) & \text{if } x_i = y_i \end{cases} \quad \text{---------- (1)}$$

$$\mathrm{Sim(X, Y)} = \sum_{\mathrm{i}=0}^{\mathrm{l}-1} \mathrm{w(X, Y, i)}. \quad (2)$$

関数 $w(X, Y, i)$ はシーケンス上におけるデータ間の部分的な一致による weight 値を算出し、Sim(X,Y) は総時間内における全体的な weight の値である。

ユーザプロファイルにコマンドシーケンスの集合である **D** を用い、そのユーザの最新のコマンドシーケンスを X とすると、その類似尺度は、

$$\mathrm{Sim_D(X)} = \max_{\mathrm{Y} \in \mathbf{D}} \mathrm{Sim(Y, X)}. \quad (3)$$

のようにして表すことができる。

この処理の特徴として、(1) 式よりイベントの一致が連続した場合に高い類似性を示すこと、また (3) 式からプロファイル全体の類似度の平均的な基準ではなく最大値を信用していることがわかる。

これらよりアルゴリズムは以下のようになる。

表 1 類似度の算出アルゴリズム

```
1.  /* 変数定義 : */
2.  while D ≠ φ do
3.      for i := 0 to l − 1 do
            (1) 式に基づく処理;
4.      (2) 式に基づく処理;
5.      od;
6.  od;
7.  (3) 式に基づく処理 (最大値を出力);
```

2.2 フリークエントエピソードの適用

フリークエントエピソード解析 [2] は、相関ルールをベースとしたアルゴリズムからなり、シーケンス構造をもつデータの頻出性の解析に有効である。

ここでエピソードとは半順序性をもって起こるイベントの集合のことであり、エピソードにはシリアルとパラレルという 2 つの種類がある。シリアルエピソードにはイベント間の順序的制約があり、パラレルエピソードに順序的な制約はない。

*A anomaly intrusion detection approach from user audit log data using frequent episode analysis

†Daisuke Ito,Hayato Ohwada

‡Faculty of Sci. and Tech., Science University of Tokyo

このような振舞いに関するルールを各ユーザごとに導出し、条件として先程のアルゴリズムに組み込む。以降にシリアルとパラレルの場合に関する処理の方法を示し、2つの処理を組み合わせた類似性評価アルゴリズムを示す。

2.2.1 シリアルエピソード

イベント E, F において、$E \Rightarrow F$ という順序制約をもつ場合、そのような順序性の一致が X 中に現れたときに高い類似度 (n) を加算することにする。定式化すると以下のようになる。

$$w(X, Y, i) = \begin{cases} 0 & \text{if i} < 0 \text{ or } x_i \neq y_i \\ 1 + w(X, Y, i-1) & \text{if } x_i = y_i \\ n + w(X, Y, i-1) & \text{if i} > 0, \\ & x_{i-1} = y_{i-1} = E \text{ and } x_i = y_i = F \end{cases} \quad \text{---------- (4)}$$

2.2.2 パラレルエピソード

イベント A, B がパラレルの関係である場合、順序性の観点から $A \Rightarrow B$、$B \Rightarrow A$ 両方のケースが成立するので、もとの $A \Rightarrow B$ の順序性の一致が X 中に見られない場合、$B \Rightarrow A$ で試す。アルゴリズムは表2のようになる。

表2 パラレルエピソードアルゴリズム

1. **if** i $> 0, (y_{i-1}, y_i) = (A, B) \text{or} (B, A)$;
2. **if** $x_{i-1} \neq y_{i-1}$ and $x_i \neq y_i$ **then**
   $tmp = y_{i-1}, y_{i-1} = y_i, y_i = tmp$;

2.2.3 アルゴリズム

類似度の算出の際に2つのエピソード処理を加える。表3はその改良したアルゴリズムである。

表3 改良した類似度算出アルゴリズム

1. /*変数定義:*/
2. **while** D $\neq \phi$ **do**
3. **for** i := 0 **to** l − 1 **do**
   /*パラレルエピソードアルゴリズム:*/
   (4) 式に基づく処理;
4. (2) 式に基づく処理;
5. **od**;
6. **od**;
7. (3) 式に基づく処理 (最大値を出力);

3 実験

実験はRedHatLinux6.2において、psacct-6.3-9のacctログを使用した。対象としたユーザは4人、データ数は15,000、シーケンスの大きさlは10、n の値を3とする。使用する15,000個のデータは10分割し、最新の1,500個から10分の1を任意に取り出しテストデータとして、残りをトレーニングデータとし9回実験を行い平均をとる。

フリークエントエピソード解析に関しては、ウインドウの大きさを5(s)、頻度が極めて高く一般的でないルールを各ユーザにつき3〜5個得た。最後に表1のアルゴリズムと、表3の改良したアルゴリズムとの類似度の値の違いを比較する。

4 実験結果

Tested	Profiled User			
User	U1	U2	U3	U4
U1	1.7 /1.7	1.8 /1.8	0.3 /0.3	1.2 /1.2
U2	1.1 /1.1	34.2/51.6	0.9 /0.9	2.1 /2.1
U3	1.3 /1.3	0.6 /0.6	55.0/75.0	1.2 /1.2
U4	0.8 /0.8	1.7 /1.7	1.2 /1.2	14.0/22.8

図1 改良したアルゴリズムとの類似度の比較
(/ の左が改良前、右が改良後の結果)

図1の結果より、U1以外のユーザは同一ユーザのテストデータで実験した場合、類似度は大きく上がった。また、異なるユーザのテストデータを与えた場合に類似度の上昇が見られないことからも、分類の精度は確実に改善されたといえる。

5 おわりに

実験により、フリークエントエピソード解析を用いてユーザの振舞いをルール化して類似度を求める本研究のアルゴリズムが有効であることを示した。現在、他のデータマイニング手法を用いた試みも行っており、それらを用いた更なる精度の向上を検討中である。

参考文献

[1] Terran Lane and Carla E. Brodley, Temporal sequence learning and data reduction for anomaly detection, ACM Transactions on Information and System Security Vol.2 No.3 , 1999.

[2] Heikki Manila, Hannu Toivonen, A.Inkeri Verkamo: Discovering frequent episode in sequences, Proc. of the 1 st International Conference on Knowlege Discovery in Databases and Data Mining, 1995.