[ポスター発表] 研究報告

セキュリティベースのデータ変換プロセスに基づく Salesforce Einstein Analytics を用いたデータ解析

輪島 幸治[1,a)]

Data Analysis using Salesforce Einstein Analytics based on a Security-based Data Transformation Process

1. はじめに

情報通信の発達により，ネットワーク接続機器を狙った分散型のサービス拒否攻撃やマルウェアも増加しており，サイバーセキュリティ分野では，数多くの製品開発や研究が行われている [1]．侵入検知システム (IDS) を対象とした研究では，挙動解析や変化点検出の分析が行われている．サイバーセキュリティ分野で分析されるデータは，IP アドレスや使用プロトコル，ポート番号など内部のセキュリティ上機微な情報を多く含む場合や，データ量と計算量コストの課題，また，オンライン上のセキュリティリスクを考慮する必要がないことから，オンプレミスで分析が行われることが多い．一方で，近年におけるコロナウイルス感染症やソーシャルディスタンスなど社会情勢の影響から，在宅勤務が必須とされた．ゆえに，オンラインでのデータ分析環境の構築が急務とされてきている．

ところで，現実社会におけるオンラインでのデータ分析環境の構築は，AI やイノベーションといった目覚ましい IT 環境の変化に対応するのと同様に，取り扱いデータに対する配慮や新規環境構築の作業コスト，およびステークホルダーとの調整など考慮事項が多いことが事実である [2]．ゆえに，本研究では，まずセキュリティ分野でオンプレミスで分析しているデータセットをデータ変換して，クラウドアプリケーションで分析することを試みた．本研究における提案システムを図 1 に示す．

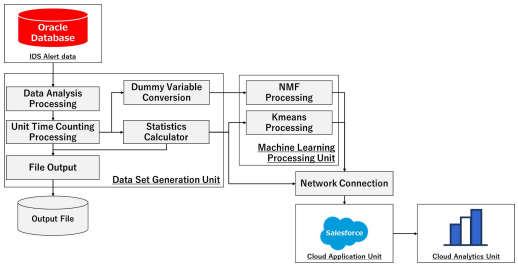

図 1　提案システムの概要

1　国立研究開発法人 情報通信研究機構
National Institute of Information and Communications Technology, Koganei, Tokyo 184–8795, Japan
a)　wajimak@nict.go.jp

提案システムは，2 種類のエンタープライズアプリケーションを用いて構築した．本研究では，Oracle Database[*1] と Salesforce[*2] を用いている．Oracle Database は，データマネジメントソリューション分野のマジック・クアドラント[*3] で，リーダーに位置付けられており，データベース管理に強みを持つ．同様に，Salesforce は，CRM カスタマーエンゲージメントセンター分野，フィールドサービス管理分野，営業支援システム分野など多くの分野で，リーダーに位置付けられているクラウドアプリケーションプラットフォームである．提案システムは，セキュリティベースのデータ変換プロセスを用いて，セキュリティ上の懸念やデータ量と計算量コストの課題を解決すること，構築済みの既存のエンタープライズアプリケーションという構築コストが少ない方法で，実装することに特徴がある．データ変換プロセスを用いることで，既存課題を解決して Salesforce Einstein Analytics [*4] でデータ解析が行える．

2. セキュリティベースのデータ変換プロセス

データ変換プロセスでは，単位時間に基づいて集計して，ダミー変数化処理と統計処理を用いて変換処理した．処理結果の概要を式 (1) および式 (2) に示す．

$$\begin{bmatrix} Time & \text{“}count\text{”} & \text{“}unique\text{”} & \text{“}freq\text{”} \\ 2020/01/01\,00:00:00 & 51 & 16 & 21 \\ 2020/01/01\,00:20:00 & 53 & 18 & 22 \\ \vdots & \vdots & \vdots & \vdots \\ 2020/12/31\,23:40:00 & 54 & 19 & 24 \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} Time & \text{“}port\text{”}\_53 & \dots & \text{“}port\text{”}\_443 \\ 2020/01/01\,00:00:00 & 1 & \dots & 1 \\ 2020/01/01\,00:20:00 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 2020/12/31\,23:40:00 & 0 & \dots & 1 \end{bmatrix} \quad (2)$$

*1　Oracle Developer : https://developer.oracle.com
*2　Salesforce Developers : https://developer.salesforce.com
*3　Gartner Magic Quadrant & Critical Capabilities
https://www.gartner.com/en/research/magic-quadrant
*4　Salesforce Einstein Analytics では，グラフによる可視化や Einstein Discovery による一般化線形モデル (GLM) などの回帰分析，予測，分類評価，異常値範囲算出などが行える．

式 (1) は “Summy Variable Conversion” における結果の例である．作成処理は各 IDS 項目で算出しており，列名の “count”，“unique”，“freq” の値はそれぞれ，集計単位時間における項目値が含まれるレコードの件数，ユニークな項目値の件数，最もレコード件数が多かった項目値のレコード件数である．式 (2) は図 1 で示した “Dummy Variable Conversion” で作成しており，ダミー変数行列と呼ぶ．本研究では，ダミー変数行列の変換処理で，スレッショルド処理を用いている．アラートデータ分類におけるダミー変数行列は，有効性が示されている [3]．本研究では，NMF を適用する観測行列に用いた．定式化したスレッショルド処理を式 (3) および式 (4) に示す．入力データにおける入力項目数は $(i = 1, ..., N)$ で，入力データは $(x_1, x_2, ...x_N)$ と定義している．項目における値のユニーク値を算出する関数を $unique(x_N)$ とした場合において，閾値を超えた場合に，分析対象としたアイテムに判定ラベルを設定した．

$$X = \left[x_1, x_2 \cdots x_N\right] \tag{3}$$

$$x_N = \begin{cases} unique(x_i) <= \text{Threshold}, & x_i \in Set \quad L0 \\ unique(x_i) > \text{Threshold}, & x_i \in Set \quad L1 \end{cases} \tag{4}$$

アラートデータを変換処理することで，数値行列データとなり，項目値を明らかにしない．また，集計処理を用いることで，データ量や計算量コストの課題も解決した．

3. 実装

本研究における評価対象は，2017 年 1 月 1 日から 2017 年 10 月 31 日までの IDS データセットであり，合計サイズは 83.3GB，総レコード数は 131,888,915 件である．1 カ月分である 8,430,462 件のデータで評価実験を行った．単位時間集計 (20 分) した際のレコード数は，2,232 件である．ゆえに，詳細分析を行う前のアラート全体の把握を目的とした，確認アラート数の削減方法としては，十分効果的である．データ変換プロセスにおける統計処理は Pandas[*5] の describe メソッド，機械学習アルゴリズムは scikit-learn で実装している．変換処理した数値行列データと Salesforce との API 連携では，salesforce-bulk [*6] を使用した．

4. 評価実験

提案システムで得られた結果を表 1 および図 2，図 3 に示す．可視化および Kmeans の対象は，2 節の式 (1) にて示した統計処理行列であり，NMF の対象は 2 節の式 (2) にて示したダミー変数行列である．表 1 が，Salesforce で回帰分析した結果である．“Lower” および “Upper” は，異常値の範囲 (下限値および上限値) であり，MAE および RMSE などは回帰分析の予測精度の指標である．結果から，Salesforce Einstein Analytics を用いることで，異常値の範囲や予測誤差などが算出できた．また，図 2，図 3 からスパイクポイントや異常値が割り当てられるクラスタなどが明らかとなり，異常値検出に応用できるのがわかる．

*5 pandas:https://pandas.pydata.org/
*6 Heroku : https://github.com/heroku

表 1 Salesforce Einstein Analytics Result(Record Count)

			ポート		IP	
	アラート数	プロトコル数	送信元	送信先	送信元	送信先
Lower	55	31	50	32	48	32
Upper	51,020	50,390	50,910	50,450	50,800	50,480
MAE	1735.0	1220.5	1245.2	1223.6	1238.5	1236.3
RMSE	5770.7	5842.0	5838.4	5845.5	5863.2	5859.4
$R\hat{2}$	0.6153	0.5993	0.5999	0.5987	0.5963	0.597

図 2 Salesforce Einstein Analytics の可視化

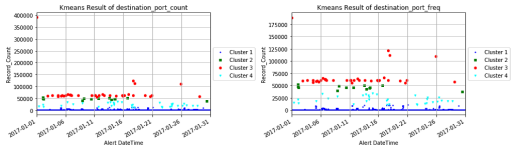

図 3 Kmeans を用いたクラスタリングの結果

NMF 評価結果は，図 2 の右図である．NMF の重み付け係数値は，分類器評価で有効性を示した [4]．本研究では，各基底の係数値の差を分布の変化とみなして異常値検出への応用を試みた．可視化結果の詳細評価や NMF の応用方法の妥当性評価は割愛して，今後の課題としたい．

5. まとめ

本研究では，セキュリティベースの変換プロセスで変換手法を用いた，アラートデータ分析システムを提案した．結果，回帰分析や可視化やクラスタリングなどで，有効な結果が得られた．データ・マイニング [5] と呼ばれる有益な知見発見手法は今後も必要とされている．Salesforce では，年 3 回のバージョンアップや，毎年の新機能発表[*7]もあり，ユーザーのニーズを取り込んだ多くの新機能がリリースされている．今後はデータの分野に限らず，データ・マイニング手法や Salesforce の新機能を活用したい．

謝辞

著者は，情報通信研究機構の IDS データセットを提供し，研究の議論の機会を与えてくれた研究者である班 涛氏に感謝したい．

参考文献

[1] M. Husák and J. Komárková and E. Bou-Harb and P. Čeleda, Survey of Attack Projection, Prediction, and Forecasting in Cyber Security,IEEE Communications Surveys Tutorials，21(1),640-660,2019
[2] マーク・ベニオフ/モニカ・ラングレー，トレイルブレイザー - 企業が本気で社会を変える 10 の思考，東洋経済新報社，2020 年 8 月 13 日
[3] 輪島 幸治，Aminanto Muhamad Erza，班 涛，伊沢 亮一，高橋 健志，井上 大介，ログのカテゴリー変数に対するダミー変数と項目マッピングを用いた行列変換処理手法，第 12 回データ工学と情報マネジメントに関するフォーラム，E1-3，2020
[4] 輪島, 幸治，非負値行列因子分解アルゴリズムに基づくメッセージ特徴の選択手法に関する研究，博士論文，筑波大学，2019
[5] 亀井 明宏，電通広告辞典，電通，2008

*7 Dreamforce : https://www.salesforce.com/dreamforce/