Vol.2021-IOT-52 No.1
2021/3/1

Salesforce Einstein Analyticsを用いたIDSデータからの脅威アラートスクリーニングの試み

輪島 幸治[1,a),b)]　高橋 健志[1,c)]　井上 大介[1,d)]

概要：近年，情報社会の発展で，あらゆる環境に情報インフラが整備され，空間・時間問わずにネットワークにアクセス可能な環境が整った．一方で，情報インフラを使用した不正通信の増加していることから，侵入検知システム (IDS) などの保護装置から得られる脅威アラートの分析が必要とされている．これまで，データ量とデータのセキュリティに関する懸念から，IDS データは，オンプレミスで分析が行われてきた．しかし，近年，在宅勤務など利便性の課題から，クラウドアプリケーションでの分析も必要とされている．そこで，本研究では，オンプレミスで行われている IDS データの分析を，IDS データを事前に加工してから，クラウドアプリケーションにて分析することにより，セキュリティに配慮しつつ異常値分析および挙動解析を行うシステムを構築した．提案システムの特徴は，2 種類の分析用データを自動生成して，クラウドアプリケーションで IDS データを分析することに特徴を持つ．提案システムを用いて，IDS データから脅威アラートをスクリーニングが行うための異常値範囲の算出や可視化をクラウドアプリケーションで行うことができた．結果を報告する．

キーワード：Salesforce，Einstein Discovery，セキュリティ，アラートスクリーニング，クラスタリング

An Attempt to Threat Alert Screening from IDS Data using Salesforce Einstein Analytics

Abstract: Recent years,information infrastructure has been improved, and an environment in which networks can be accessed regardless of space or time has been established. On the other hand, the number of unauthorized communications that abuse information infrastructure is increasing. Therefore, it is essential to analyze threat alerts obtained from protection devices such as intrusion detection systems (IDS). Currenty,IDS data analysis has been performed on-premises due to concerns about data volume and data security. However, due to the issue of convenience, analysis in cloud applications is required. Therefore, in the proposed system of this research, visualization of IDS data and detection of outlier range are analyzed by cloud applications. In this research, in order to consider security, the analysis of IDS data performed on-premises is characterized by processing the IDS data in advance and then analyzing it with a cloud application. The feature of the proposed system is that the item statistical matrix and the dummy variable matrix are converted and the analysis data is automatically generated.In addition, one of the features is to analyze the created data set with a cloud application. We were able to calculate and visualize the outlier range with a cloud application for screening threat alerts from IDS data using the proposed system. Report the results.

Keywords: Salesforce，Einstein Discovery，Security，Alert Screening，Clustering

1. はじめに

近年, 情報システムのセキュリティ向上を目的に, 多くの脅威アラート分析などのセキュリティ対策が提案されている．不正通信分析では，侵入検知システム (IDS:Intrusion Detection System) に格納された大量の IDS データを分析

[1] 国立研究開発法人 情報通信研究機構
National Institute of Information and Communications Technology
4-2-1, Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan
[a)] wajimak@nict.go.jp
[b)] kwajima@ce.slis.tsukuba.ac.jp
[c)] takeshi_takahashi@nict.go.jp
[d)] dai@nict.go.jp

しなければならない．これまで，クラウドアプリケーションで IDS データを分析することは，セキュリティ上の懸念や，取り扱いデータ量から困難とされてきた．また，IDS データは，組織内部の通信情報を含む機微なデータであり，クラウドアプリケーションで機微なデータを管理することが，セキュリティポリシ違反となる組織も少なくない．しかし，近年，在宅勤務が必要とされる状況の増加や，システムの運用管理において，クラウドアプリケーションがますます必要とされるようになってきた．そこで，本研究では提案システムで，この課題の解決を試みる．

本論文では，文献 [1] にて提案した，提案システムについて，詳述する．提案システムを用いることで，オンプレミスで管理されている IDS データをセキュリティポリシに配慮しつつクラウドアプリケーションで分析することができる．クラウドアプリケーションは，多くの種類があるが，本研究における提案システムには，米国セールスフォース・ドットコム（日本法人：株式会社セールスフォース・ドットコム）[*1]が，プラットフォームとして提供している Salesforce を Web アプリケーション基盤に用いる．

Salesforce を Web アプリケーション基盤に用いることで，時系列データを可視化することや，モデリングアルゴリズムを適用できる．提案システムは，大きく 4 種類のユニットから構成されている．まず，IDS データを事前に加工処理して，2 種類の分析用データセットを作成する “Data Set Generation Unit”．次に，特徴変換アルゴリズム及びクラスタリングアルゴリズムを適用する “Machine Learning Processing Unit”．そして, 2 つのクラウドアプリケーションである，“Cloud Application Unit” 及び “Cloud Analytics Unit” である．“Cloud Application Unit” 及び “Cloud Analytics Unit” では，時系列可視化，単位時間の基底重要度，クラスタリング，回帰分析で評価を行う．

提案システムの特徴は，IDS データの変換処理を行うことから，クラウドアプリケーション上に機微なデータは，保持しないという特徴を有している．また，単位時間でのレコード集計をしていることも特徴の一つである．集計処理することで，作成データセットのサイズが減少することから，提案システムでは，データサイズを課題としない．ゆえに，大量データである場合でも，クラウドアプリケーションで分析が可能となる．加えて，Salesforce Einstein Analytics を用いることで，平易に可視化や異常値の範囲の算出，回帰分析が行える．

本論文では，2 章で関連研究に関して述べる．3 章で，提案システム概要を示して，Salesforce Einstein Analytics で，IDS データを分析する方法を示す．4 章では，実験と評価対象について述べ，5 章及び 6 章で実験結果と考察を行う．最後に 7 章で，まとめと今後の課題を示す．

[*1] Salesforce - セールスフォース・ドットコム: https://www.salesforce.com

2. 従来研究

2.1 侵入検知と異常検知に関する研究

脅威アラートには，TCP SYN Flood のような大量のパケットを送信して，通信のコネクション確立要求を行う DoS 攻撃 [2] などがあり，侵入検知や異常検知などの検出を目的に，数多くの研究が行われている [3][4]．従来研究における脅威アラートの検出方法は，ミスユース検出と異常検知に大別されている [5]．ミスユース検出は，不正アクセスのパターンが登録されているデータベースと照合して判別する方法であり [5]，異常検知は，以前は未知であった攻撃パターンを明らかにする方法である．既存研究 [5] における異常の定義は，正常なパターンに該当しないパターンであると定義している．

本研究では，IDS アラートデータからの異常検知を目的としている．IDS アラートデータから異常を検出する典型的な方法としては，外れ値検出，変化点検出，異常値部位検出などがあり [6]，所定した閾値を越えたら異常と判断する定義としてのネイマン・ピアソン決定則がある [6]．また，正規分布や多項分布，経験分布など分布で，異常値を判断するベイズ決定測 [6] や，マージン最大化近傍法などの手法もある [6]．

第 3 節にて後述するが，提案システムでは，異常検出を目的に，時系列分析を用いた変化点検出，Kmeans を用いた外れ値の検出，NMF を用いた異常値部位検出を試みる．

2.2 IDS のベンチマークデータセットに関する研究

IDS は，異常検知など通信を監視するシステムである．IDS におけるアラート分析は，商用 IDS システムにおける分析として，以前から行われてきた [7][8]．研究においては，IDS システムの攻撃分類や事前スキャンなど，多くの研究が行われている [9][10]．一般に，IDS データは，組織内部の通信情報を含む機微なデータとして取り扱われている．ゆえに，IDS に関する研究には，研究用のデータが用意されており，UNSW-NB15[11] などが，ベンチマーク用のデータとして提供されている．具体的に，UNSW-NB15 を用いた研究には，攻撃検出性能の比較，新モデルを用いた性能評価，比較など多くの研究が行われている [12][13][14]．

ところで，IDS データは一般にオンプレミスで分析が行われている．これは，クラウドアプリケーションで，IDS データを取り扱うことを，セキュリティポリシ違反とする組織も少なくないためである．しかし，近年では，在宅勤務などで，クラウドアプリケーションが必要とされる状況も増えてきた．このため，IDS データをクラウドアプリケーションでデータ管理を行い，可視化，機械学習などで分析する試みは，利便性の観点で重要な課題である．本研究では，IDS データに対して，セキュリティベースの変換処理を行い，機微なデータとして取り扱われているデータを，クラウドアプリケーション上で分析することを試みた．

3. 提案法

本研究では，Salesforce Einstein Analytics を用いて，クラウドアプリケーション上で，IDS データから脅威アラートのスクリーニングを試みる．本研究の提案である提案システムの概要を，図 3 に示す．

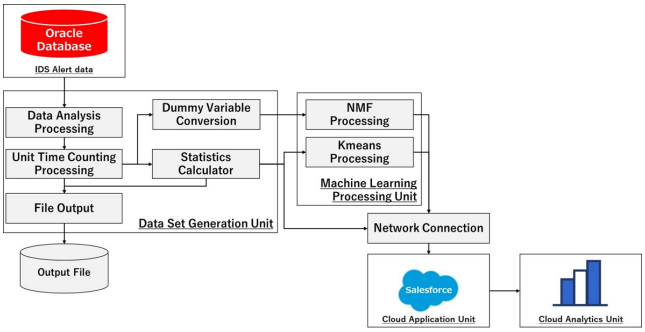

図 1 提案システムの概要

図 3 の提案システムでは，オンプレミスのデータベースに格納されているデータを，処理ステップを介して，クラウドアプリケーションに格納可能なデータに変換している．提案システムにおける処理ステップ及び各ユニットの概要を下記に示す．

Step 1. Data Set Generation Unit
項目値の単位時間集計を行い，ダミー変数化処理及び項目統計処理を用いて，時系列分析を行う分析データセットを自動作成する．第 3.1 節にて示す．

Step 2. Machine Learning Processing Unit
自動生成した分析用データセットから，機械学習アルゴリズムを用いて，機械学習処理を行う．第 3.2 節にて示す．

Step 3. Cloud Application Unit
自動生成されたデータセット及び機械学習処理された結果を，Network Connection を介して，クラウドアプリケーションに連携する．第 3.3 節にて示す．

Step 4. Cloud Analytics Unit
クラウドアプリケーションに連携したデータに対して，可視化及びモデリングアルゴリズムを使用した分析を行う．第 3.4 節にて示す．

本研究では，上記の *Step*1 から *Step*2 を，オンプレミス側で処理することで，クラウドアプリケーション上での IDS データの分析に可能とした．

3.1 Data Set Generation Unit

提案システムで分析データセットを自動作成するユニットを説明する．本研究では，下記に示す 2 種類の分析データセットを作成する．

(1) ダミー変数行列 (result of dummy variable conversion)
(2) 項目統計行列 (result of a statistics calculator)

ダミー変数行列は，次元削減を用いたアラート検出に用いる．項目統計行列は，時系列解析とクラスタリングに用いられる．本研究では，自動作成する分析データセットは，IDS データを単位時間ごとに集計して作成される．集計する単位時間が 20 分である場合の例を式 (1) 及び式 (4) に示す．まず，ダミー変数行列と行列における閾値処理について，式 (1) から式 (3) に示す．

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

式 (1) のダミー変数行列は，ダミー変数化処理を用いて，IDS 項目値のカテゴリー変数値を，数値行列データに変換することで作成される．ダミー変数化処理は IDS 項目ごとに実行され，式 (1) は，ダミー変数化処理をポート番号に対して，適用した場合における結果の例である．一方で，ダミー変数化処理は，IDS 項目の値に一意の値が多い場合，計算量の課題で，変換できない場合がある．そこで，本研究では変換処理における閾値を設定した．変換処理を閾値で判断する箇所を定式化した表現を式 (2) 及び式 (3) に示す．

$$X = \left[x_1, x_2 \cdots x_N\right] \tag{2}$$

$$x_N = \begin{cases} unique(x_i) <= \text{Threshold}, & x_i \in Set \quad L0 \\ unique(x_i) > \text{Threshold}, & x_i \in Set \quad L1 \end{cases} \tag{3}$$

式 (2) では，入力データにおける入力項目数は $(i = 1, ..., N)$ で，入力データを $(x_1, x_2, ...x_N)$ と定義した．また，式 (3) における判定ラベルは ($L0$ もしくは $L1$) と定義した．項目における値のユニーク値を算出する関数を $unique(x_N)$ と表している．したがって，本研究における変換処理では，式 (2) 及び式 (3) から，入力データにおける入力項目のユニーク値が，閾値を超えた場合，ダミー変数化処理を行わない判定ラベルを設定した．

これまでの研究で，ダミー変数化処理する際に，処理の実現可否を，閾値で判断する処理を追加した場合の有効性が示されている [16]．式 (3) における閾値は，計算機のメモリ容量で決定される．ダミー変数化処理では，判定ラベルが $L0$ と設定した IDS 項目が，変換対象とされ第 3.3 節を介して，Salesforce に格納して，評価に使用する．

次に，項目統計行列を式 (4) に示す．式 (4) の項目統計行列は，IDS 項目ごとに “count”，“unique” 及び “freq” の 3 種類の値を計算して作成する．

$$\begin{bmatrix} Time & \text{“}count\text{”} & \text{“}unique\text{”} & \text{“}freq\text{”} \\ 2020/01/01\,00:00:00 & 51 & 16 & 21 \\ 2020/01/01\,00:20:00 & 53 & 18 & 22 \\ 2020/01/01\,00:40:00 & 55 & 20 & 23 \\ : & : & : & : \\ 2020/12/31\,23:40:00 & 54 & 19 & 24 \end{bmatrix} \tag{4}$$

“count” 項目は，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である．“unique” 項目は集計単位時間における項目値のユニーク件数である．“freq” 項目は，集計単位時間における項目値のうち，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．本研究で分析対象とする IDS 項目は，4 章にて，後述する．

3.2 Machine Learning Processing Unit

自動生成したデータセットに機械学習アルゴリズムを適用するユニットを説明する．本研究における機械学習アルゴリズムでは，特徴量変換アルゴリズム及びクラスタリングアルゴリズムを用いる．データセットは，第 3.1 節にて示した式 (1) のダミー変数行列及び式 (4) の項目統計行列から作成データセットを用いる．評価方法及びアルゴリズムと対象行列の組み合わせを表 1 に示す．

表 1 アルゴリズムと対象行列の組み合わせ

評価方法	クラスタリング	重要度算出
アルゴリズム	Non-Negative Matrix Factorization(NMF)	Mini Batch K-Means
対象行列	式 (1) ダミー変数行列	式 (4) 項目統計行列

本研究における，特徴量変換アルゴリズムは，非負値行列因子分解 (Non-Negative Matrix Factorization : NMF) を採用しており，変換特徴量を重要度算出で用いた．本研究における NMF の位置付けだが，本研究では，NMF は異常値検出するための，アラートの重要度を算出及び可視化することを目的としている．しかし，アラートデータは，定量的な重要度を出すことが困難な場合が多い．一方で，NMF の基底における重み付け係数値は，分類器における評価で有効性が示されている [15]．ゆえに，重み付け係数値を重要度とみなし，項目値の分布の変化など，重要度に変化が現れた場合に，その係数値から基底への適合度を評価して，アラート検出への適用を試みた．

クラスタリングアルゴリズムには，Mini Batch KMeans を採用した．K-means の評価対象行列は項目統計行列である．Kmeans は古典的なアルゴリズムであるが，Mini Batch KMeans は，Kmeans のバリエーションの一つである．本研究では，データに異常に大きな値が含まれている場合，その値を持つ項目は別のクラスタに割り当てられることから，本研究では，クラスタリングを異常値検出の評価に用いた．Mini Batch KMeans は，計算時間を短縮が行えるなどのメリットがある．

3.3 Cloud Application Unit

本研究では，運用管理など実務上の利便性から，クラウドアプリケーションに，Salesforce を用いる．本研究における，Salesforce 実装内容の概要を下記に示す．

提案システムにおける Salesforce 設定の概要

(1) User Interface : Lightning Experience
(2) Database Table : Custom Object
(3) Network Connection : Monitoring Async Api Jobs
(4) Implementation : salesforce-bulk library
(5) Salesforce Edition : Developer Edition

Salesforce は，クラウドアプリケーションだが，クラウド型のデータベースとして，データベースのように各種データを管理できる．また，インターフェースをフラットデザイン化すること，REST API , SOAP API など任意の連携方法を指定して，API 連携することもでき，中長期的なデータの運用管理にも強みを持つ．各機能の詳細については，開発者ドキュメントを参考にされたい[*2]．

3.4 Cloud Analytics Unit

本研究では，提案手法における Cloud Analytics Unit に Salesforce Einstein Analytics を使用した．Salesforce Einstein Analytics は，標準機能では有効となっていないことから，評価実験では，Einstein Analytics Plus ライセンスが有効となっている Salesforce 組織を用いている．

本研究では，第 3.1 節及び第 3.2 節で得られた結果を Salesforce に格納して可視化することで，実務上の利便性に対応した．アラートスクリーニングの方法だが，本研究における主な分析は，作成データセット及び機械学習アルゴリズムの結果を Salesforce Einstein Analytics で可視化を試みた．加えて，本研究では，Einstein Discovery を用いた回帰分析でも分析を試みる．回帰分析で使用した Einstein Discovery 分析の設定内容は，下記の通りである．

回帰分析を行う Einstein Discovery の設定

(1) Story Goal :Minimize
(2) Story Type :Insights & Predictions
(3) Story Settings :Correlation Top 10 Item Selection
(4) Algorithm :Generalized linear model(GLM)

4. 実装

4.1 Experiments Environment

本研究では, Python と Salesforce で実装して, 統計処理はライブラリの “pandas”，機械学習アルゴリズムはライブラリの “scikit-learn” で実装した．本研究における Salesforce との Network Connection の実装だが，Heroku が github で公開しているライブラリの salesforce-bulk[*3]を介した連携で，データの一括読み込みジョブ機能を使用している．

[*2] Salesforce Developers : https://developer.salesforce.com
[*3] heroku - salesforce-bulk : https://github.com/heroku/salesforce-bulk

既存ライブラリを介すことで，詳細な設定を意識することなく，平易に Salesforce と連携が行える．Salesforce のエディションは，Developer Edition を用いており，評価実験を行った時点の Salesforce のバージョンは，Salesforce Summer '20 である．本研究の評価対象として，情報通信研究機構の LAN で 2017 年 1 月 1 日から 2017 年 10 月 31 日までの 10 ヶ月間に発報された IDS データを用いた．合計データサイズは 83.3GB で，合計数は 131,888,915 件である．本研究では，そこから 1 カ月分に相当する 2017 年 1 月のデータを抽出して，評価実験を行った．脅威アラートと判断する箇所だが，正解データで評価する方法はなく，可視化及びアルゴリズムを用いて，IDS データ状況から，異常値が出現する状況を脅威アラートが発生した時点として判断した．

また，本研究では，提案手法を用いた評価実験で，IDS 項目を選択している．IDS 項目の選択項目は，アラート数，プロトコル数，送信元ポート数，送信先ポート数，送信元 IP 数，送信先 IP 数を使用した．また，ダミー変数化処理で用いる閾値の値には，20,000 を設定した．本研究では，IDS データの管理を，Oracle Database で行った．[*4] Oracle Database のバージョンは，Oracle Database19c である．

4.2 Salesforce Einstein Analytics を用いた評価方法

本研究では, Salesforce Einstein Analytics を用いて, IDS アラートデータから，脅威アラートの評価を行う．本研究における脅威アラートの評価方法を下記に示す．

脅威アラートの評価方法

(1) 時系列データ可視化 - 統計行列からの異常値検出
(2) NMF 可視化 - 観測行列からの分布異常検出
(3) Kmeans クラスタリング - Kmeans 結果の可視化
(4) 脅威アラート判別用の異常値範囲の算出

(1)，(3) 及び (4) は，項目統計行列を用いた評価である．(1) は，Salesforce Einstein Analytics で可視化する．(3) は，Kmeans を適用して，Salesforce 上にて可視化する．(4) は，Salesforce Einstein Analytics の Einstein Discovery で算出する．(2) は，ダミー変数化行列に対して，NMF を適用して，係数行列に変換して，特性を表す基底を可視化する可視化分析である．特性を表す基底が明らかとなれば良いが，ダミー変数行列の大きさから，NMF は 20 分ごとに適用して，算出した結果を使用している．ゆえに，同一次元かつ同一基底であるからといって，算出される値は推移で使用できない．したがって，分布の異常性の判断のみで使用した．NMF は，適用対象における，観測行列の生成プロセスに応じて，乖離度基準が異なる．そこで，本研究では，NMF を 20 分ごとに適用していることから，観測行列の生成プロセスに変化があった場合に，日時において，異常な状態であると判断して，異常値検出に適用した．

[*4] Oracle Developer : https://developer.oracle.com/databases/

5. 評価 - Salesforce Einstein Analytics

本研究では，アラートスクリーニングを行う．1 か月間の総抽出データ数は 8,430,462 である．提案した方法を用いて，20 分間で単位時間集約の場合，レコードの総数に関係なく 31 日で 2232 となる．したがって，提案手法の結果，8,430,462 件のレコードが，2232 件のレコードに集約されることとなる．ゆえに，アラート数で，異常値が多数出現するような日時を，脅威アラートがあったとして日時を特定するのを目的とする場合，確認アラート数が 90 % 削減できることに相当する．ゆえに，提案システムを，スクリーニングにも用いることができると言える．

本章では，Salesforce Einstein Analytics を用いて，大量の IDS データを対象に，アラートデータ処理を改善した結果を示す．重要なアラートデータの判断方法においては，表層情報であるスパイクポイント，クラスタリング，重要度評価である．

5.1 時系列データ 可視化

Salesforce Einstein Analytics を用いて，アラートデータを可視化した結果を図 2 から図 7 を示す．図 2 から図 7 の可視化結果は，第 3.1 節における式 4 の項目統計行列を用いている．図 2 はアラート件数の可視化結果である．同様に，図 3 は，プロトコル項目，図 4 は，送信元 IP 項目，図 5 は，送信元ポート項目，図 6 は，送信先 IP 項目，図 7 は，送信先ポート項目の可視化結果である．

図 2 から図 7 における縦軸は単位時間中のレコード件数であり，横軸は，アラートが発報された日時である．図 2 は，アラート件数のみの可視化結果である．図 3 から図 7 については，項目の各統計行列が可視化されており，上段がユニークな項目のレコード件数 (unique)，中段が最頻値項目のレコード件数，下段が項目のレコード件数 (count) を表している．

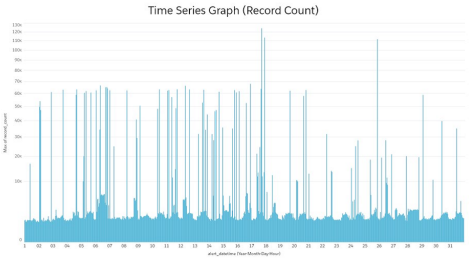

図 2 Time Series of Record Count

図 2 から図 7 から明らかなように，時系列におけるスパイクポイントが可視化されていることがわかる．結果から，時系列データ分析によって，IDS データにおける定常性と非定常性を明らかとなったことから，脅威アラートをスクリーニングするための異常値と傾向が明らかとなった．各グラフにおける個別の考察結果については割愛する．

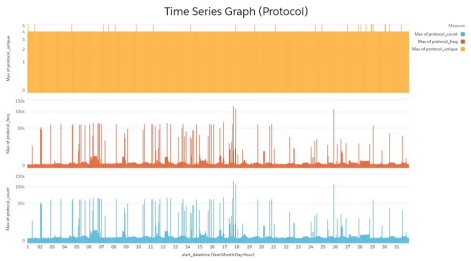

図 3 Time Series of Record Count

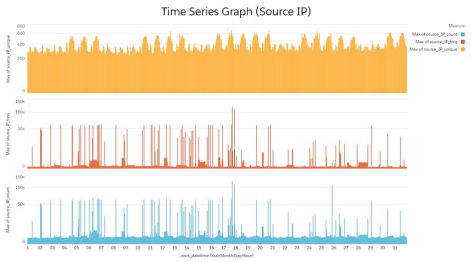

図 4 Time Series of Source IP

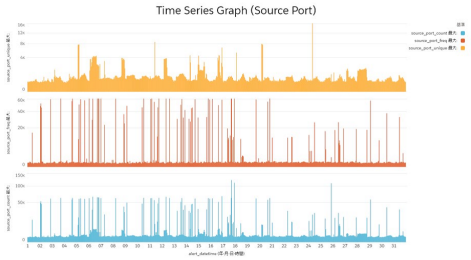

図 5 Time Series of Source Port

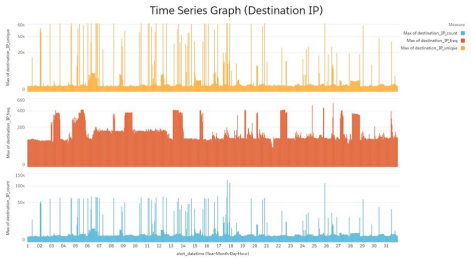

図 6 Time Series of Destination IP

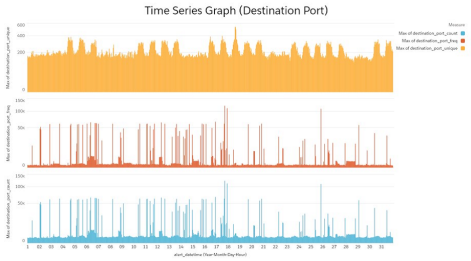

図 7 Time Series of Destination Port

5.2 NMF 可視化

次に，オンプレミスで NMF アルゴリズムを適用して，Salesforce で，可視化した結果を図 8 から図 10 に示す．NMF での評価は，ダミー変数行列を次元削減して，各基底で係数値の合計を算出した結果を用いている．図 8 は，各日での集計結果である．図 8 の左側がアプライアンスでの評価，図 8 の右側はプロトコルでの評価結果である．縦軸は日，横軸は各基底の合計値を算出して，一方の基底からもう一方の基底の合計値を減算した結果である．図 9 及び図 10 は，1 ヶ月間を 20 分ごとに可視化した結果である．図 9 はアプライアンスでの評価，図 10 はプロトコルでの評価結果である．縦軸は図 8 における横軸と同様で，各基底の合計値を算出して，一方の基底の合計値から，もう一方の基底の合計値を減算した結果である．横軸は日時である．

NMF での評価は，観測行列の生成プロセスに変化があった場合の検知に用いている．値が 0 に近い場合に異常値と判定している．結果は，0 に近い数値がいくつか観測することができたことから，観測行列の生成プロセスに変化がある日を検知することができた．

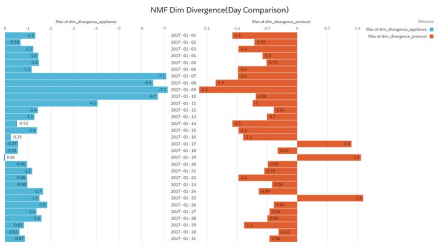

図 8 NMF Result of Day Comparison

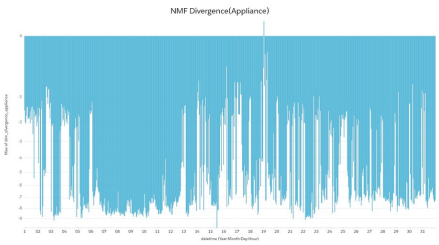

図 9 NMF Result of Appliance

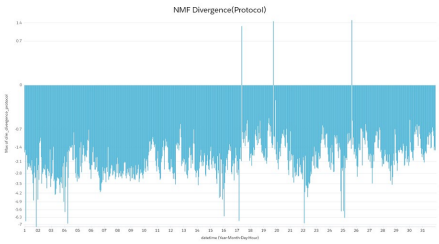

図 10 NMF Result of Protocol

6. 評価 - Kmeans クラスタリング

NMF と同様に，オンプレミスで Kmeans を適用して，Salesforce にデータ連携した Kmeans でのクラスタリング結果を図 11 から図 14 に示す．本研究では，送信元ポート及び送信先ポートの単位時間におけるレコード件数 (count)，最頻値項目のレコード件数を評価対象としている．図 11 は送信先ポートのレコード件数，図 12 は送信先ポートの最頻値項目のレコード件数，図 13 は送信元ポートのレコード件数，図 14 は送信元ポートの最頻値項目のレコード件数である．図 11 から図 14 における縦軸は単位時間中の算出値であり，横軸は，アラートが発報された日時である．

クラスタリング結果だが，図 11 から図 14 から明らかなように，“Cluster 3” に異常値が割り当てられており，クラスタリングで異常値検出が行えていることがわかる．各グラフにおける個別の考察結果については割愛する．

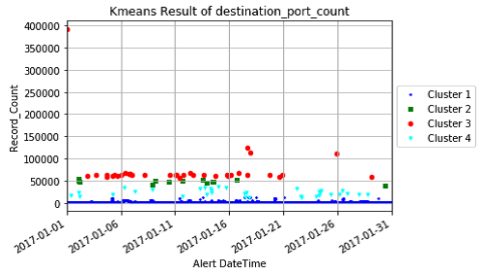

図 11 Kmeans of destination port (count)

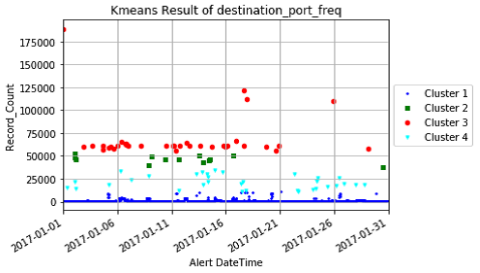

図 12 Kmeans of destination port (freq)

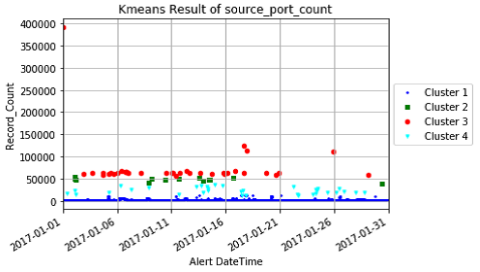

図 13 Kmeans of source port (count)

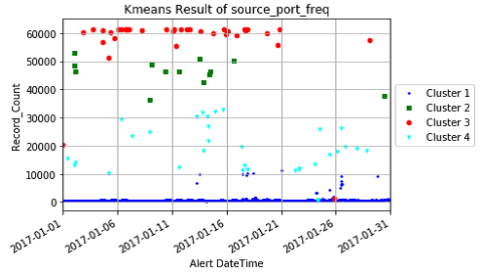

図 14 Kmeans of source port (freq)

7. 評価 - Einstein Discovery

提案システムにおける Einstein Discovery の評価だが，Salesforce の Einstein Discovery は，多くの機能を有する[*5]．本研究では，Einstein Discovery で，異常値範囲算出及び回帰分析を行いて分析を行う．回帰分析では，脅威アラートとしたい変数の項目値を，目標変数に用いることで，IDS データの項目値が，最小となる場合のプロトコルやポート番号などの説明変数の影響を評価できる．ゆえに，本研究では，回帰分析で脅威アラートの説明変数の影響を評価できるとみなして，目標変数に対する相関係数が高い上位 10 変数を入力変数に用いて回帰分析で評価した．

本研究における Einstein Discovery での回帰分析モデルには，一般化された線形モデル (GLM) を用いている．また，Einstein Discovery で評価する場合の入力値には，第 3.1 節の式 (4) にて示した項目統計行列にて算出した各集計値を，用いている．なお，本研究における回帰分析の評価指標値には，目標変数に対する予測誤差には，MSE，RMSE，MAE を用いており，回帰分析モデルの評価には，R^2 及び AIC を用いている．まず，回帰分析モデルで，目標変数を予測した場合の予測誤差の結果を表 2 に示す．

表 2 List of GLM Prediction Error Results

No.	Item Name	MSE	RMSE	MAE
1	Record Count	33403329.44	5770.6549	1734.9961
2	Protocol Count	34199420.74	5841.9721	1220.5034
3	Protocol Freq	34920580.93	5900.6848	1222.6577
4	Protocol Unique	**0.0248**	**0.1574**	**0.0345**
5	Source IP Count	34443096.8	5863.241	1238.5398
6	Source IP Freq	52885685.58	7213.4207	1867.088
7	Source IP Unique	**1262.5093**	**35.5132**	**26.9235**
8	Source Port Count	34147036.8	5838.4361	1245.2302
9	Source Port Freq	51187811.37	7081.3119	1904.2308
10	Source Port Unique	**623449.228**	**774.2855**	**229.0845**
11	Destination IP Count	34395426.67	5859.416	1236.2735
12	Destination IP Freq	**1674.3328**	**40.8974**	**24.6805**
13	Destination IP Unique	24877401.2	4959.7104	1135.9258
14	Destination Port Count	34239233.82	5845.4957	1223.6159
15	Destination Port Freq	34573648.16	5870.0788	1217.0179
16	Destination Port Unique	**816.0953**	**28.5578**	**21.4796**

結果，“No.4” であるプロトコルのユニーク数において，有効性を示すことができた．加えて，“No.7” である送信元 IP のユニーク数，“No.10” である送信元ポートのユニーク数，“No.12” である送信先 IP の最頻値レコード件数，“No.16” である送信先ポートのユニーク数などでも，相対的に，目標変数との誤差が少ないことが明らかとなった．したがって，これらの目標変数で用いた説明変数は，脅威アラートに対しての影響を評価できる説明変数とみなせる．個別の説明変数の評価や詳細については，割愛する．

[*5] Einstein Discovery での説明、予測、アクション: https://help.salesforce.com/articleView?id=bi_edd.htm

次に，異常値範囲算出と回帰分析を表 3 示す．

表 3 List of GLM Model Evaluation Results

No.	Item Name Range	Non Outlier	R^2	AIC
1	Record Count	55 - 51,020	**0.6153**	**11254.8597**
2	Protocol Count	31 - 50,390	0.5993	582754.9477
3	Protocol Freq	14 - 49,650	0.5847	629296.0062
4	Protocol Unique	3.207 - 4.791	0.0076	**4363.7671**
5	Source IP Count	50 - 50,910	0.5963	576903.0624
6	Source IP Freq	4 - 47,050	0.3328	1356492.506
7	Source IP Unique	-	**0.862**	**6001.2395**
8	Source Port Count	32 - 50,450	0.5999	583310.0516
9	Source Port Freq	7 - 44,070	0.2572	1708604.324
10	Source Port Freq	20 - 6,163	0.4682	120499.6511
11	Destination IP Count	48 - 50,800	0.597	577593.9073
12	Destination IP Freq	-	**0.7328**	**11434.8733**
13	Destination IP Unique	34 - 45,430	**0.6485**	504956.7761
14	Destination Port Count	32 - 50,480	0.5987	582290.1856
15	Destination Port Freq	6 - 49,280	0.5896	634706.1751
16	Destination Port Unique	19 - 523.8	**0.8011**	**5662.9945**

結果，異常値範囲においては，各項目値の異常値の範囲が明らかとなった．ゆえに，各項目で算出された値から，異常値を判別でき，脅威アラートのスクリーニングが行える．また，“No.1”，“No.7”，“No.12”，“No.13”，“No.16” などといったいくつかの目標変数での回帰分析で，R^2 及び AIC の評価指標で，回帰分析モデルの有効性が明らかとなった．

8. まとめ

本研究では，通常，オンプレミスで分析・管理されている IDS データを，クラウドアプリケーションにて，脅威アラートを分析するための変換プロセスを持つシステムを提案した．統計行列データセットとダミー変数データセットという 2 種類の分析データセットを作成して，Salesforce Einstein Analytics にて，異常値の可視化，単位時間における基底重要度，クラスタリングの可視化，目標変数の予測誤差，回帰分析モデルの評価を行った．結果，ネットワーク状況から，脅威アラートと判断するために必要な項目値の異常値を判別する範囲を明らかとすることができた．また，時系列グラフでの可視化，脅威アラートを評価するための傾向として，IDS データにおける定常性と非定常性を明らかにすることができた．加えて，Kmeans を用いたクラスタリングで，明確な有効性を確認することができた．そして，NMF を用いた観測行列の分布に変化がある場合の異常値検出法，回帰分析における評価でも，一定の有効性を明らかとすることができたと言える．

今後の課題は，単月だけではなく，複数月の IDS データを用いた評価を行うこと．ネットワークの通信状況における定常性と非定常性を明らかにして，評価を行うこと．

そして，本研究で得られた結果から，脅威アラートを定義して，IDS データから，適切にスクリーニングすることで，セキュリティを向上させることを目標としたい．

参考文献

[1] 輪島 幸治, 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, インターネットと運用技術シンポジウム論文集, 95–96, 2020, nov 2020.

[2] 寺田 真敏, DoS 攻撃：1. DoS/DDoS 攻撃とは, 情報処理, No. 5, Vol. 54, 428-435, Apr 2013.

[3] 伊波 靖, 高良 富夫. サポートベクタマシンを用いた WAF への異常検知機能の実装と評価. 情報処理学会論文誌コンピューティングシステム（ACS）, 1, 7, 1–13, mar 2014.

[4] I. Butun and P. Österberg and H. Song. Security of the Internet of Things: Vulnerabilities, Attacks, and Countermeasures. IEEE Communications Surveys &Tutorials, 1, 22, 616-644, 2020.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. ACM Comput. Surv., No. 3, Vol. 41, July 2009.

[6] 井手 剛，杉山 将，機械学習プロフェッショナルシリーズ - 異常検知と変化検知，講談社，2015.

[7] 神鳥 泰章, Part2 ブロードバンド時代のセキュリティ対策 (5)IDS 技術とその動向 (セキュリティ最新動向) , Business communication, ,No. 12, Vol. 39, 111–114, Dec 2002.

[8] 高橋 正和, セキュリティ最新動向 (6) 不正侵入検知装置 (IDS) の概要と最新動向について〔含 News&Topics セキュリティ関連製品&サービスの最新情報〕, Business communication, ,No. 6, Vol. 41, 93–98, Jun 2004.

[9] L. N. Tidjon and M. Frappier and A. Mammar. Intrusion Detection Systems: A Cross-Domain Overview. IEEE Communications Surveys &Tutorials, 4, 21, 3639–3681, 2019

[10] Ankush Singla, Elisa Bertino, Dinesh Verma. Preparing Network Intrusion Detection Deep Learning Models with Minimal Data Using Adversarial Domain Adaptation. Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, 127-140, Oct 2020.

[11] N. Moustafa and J. Slay, UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set), 2015 Military Communications and Information Systems Conference (MilCIS), 1-6, 2015.

[12] A. Binbusayyis and T. Vaiyapuri. Identifying and Benchmarking Key Features for Cyber Intrusion Detection: An Ensemble Approach. IEEE Access, No. 7, July 2019.

[13] K. Jiang and W. Wang and A. Wang and H. Wu. Network Intrusion Detection Combined Hybrid Sampling With Deep Hierarchical Network. IEEE Access, Vol. 8, 2020.

[14] B. A. Tama and L. Nkenyereye and S. M. R. Islam and K. Kwak. An Enhanced Anomaly Detection in Web Traffic Using a Stack of Classifier Ensemble. IEEE Access, Vol. 8, 2020.

[15] 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

[16] 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（第 18 回日本データベース学会年次大会）, E1-3, 2020