Computer Security Symposium 2020
26 - 29 October 2020

Web データの回帰分析によるセキュリティ評価手法

森 拓海 [1,*]　藤田 真浩 [1]　山中 忠和 [1]

概要：サプライチェーンにおいて，取引先企業や製品の仕向け地などによって様々なセキュリティ上のリスクがある．このようなサプライチェーンのセキュリティ評価は，過去の取引実績や知名度などの主観に頼ることが多く，セキュリティ水準の低い取引先や仕向け地を選択することで，セキュリティ事故やサイバー攻撃の対象となるリスクが高まる．筆者らはこれまでに，複数の情報源から収集したオープンデータを利用して，セキュリティ水準を数値化するサプライヤー企業のセキュリティ評価手法を検討し，サプライチェーン上のサプライヤー評価方法の枠組みを示した．本稿では前述のサプライヤー評価方法の具体化に言及し，評価対象のサプライヤー企業に関するセキュリティニュースからクラスタリング分析により抽出したキーワード群を SNS(Social Networking Service)や Web 検索エンジンの表示順序を利用して評価し，選定されたキーワード群を説明変数とした重回帰分析によりセキュリティ評価値を算出する方法を提案する．また，本方式の実装検討を実施し，評価値算出のための重回帰式の精度向上のためのチューニングポイントが，クラスタリングによる説明変数候補の抽出と，説明変数候補からの独立性の高い候補の選択であること明らかにした．

キーワード：サプライチェーンセキュリティ，Web，回帰分析，セキュリティ評価手法

Security Rating Method by Regression Analysis of Web Data

Takumi Mori [1,*]　Masahiro Fujita [1]　Tadakazu Yamanaka [1]

Abstract: On the supply chain, there are various security risks caused by factors such as the business partners and the destinations of products. Supply chain security ratings are often calculated subjectively base on the past transaction records and popularity of company names. If companies chose the suppliers and destinations for products with low security levels, the risks of security incidents and cyber-attacks against products is increased. To calculate the ratings objectively, we have proposed a security rating method of supplier companies, which quantifies the security level by using open data collected from multiple information sources. We have also proposed a framework in which the supplier rating method is incorporated. In this paper, as a concrete instance of the method, we proposed a method using multiple regression analysis. The method uses keyword groups as explanatory variables. The keyword groups are selected as follows: (1) collecting security news about supplier companies, (2) extracting the keywords from the news by clustering analysis, and (3) evaluating the ranks of the keywords based on SNS or Web search engine rank. As a result of implementation of the method, we found two tuning points for improving the accuracy of the method: lining up candidates of explanatory variables by clustering, and selecting highly independent values from these candidates.

Keywords: Supply chain security, Web, Regression Analysis, Security rating method

1. はじめに

IoT の普及により，１つの製品・サービスを開発するには，サプライヤーの協力が不可欠であり，安全・安心なサプライチェーンの形成が求められる．カスペルスキーは，2017 年に発表した脅威予測レポート[1]の中で，サプライチェーン攻撃の増加を指摘し，ソフトウェア製品の正規のアップデートに ShadowPad, ExPetr/NotPetya などのマルウェアが混入した事例を紹介している．いずれの事例も，サプライチェーン上のセキュリティ水準の低い組織を狙って攻撃したものである．カスペルスキーは 2020 年度のレポート[2]において，このようなサプライチェーン攻撃を，対処が非常に難しい攻撃方法として続くことを指摘している．

サプライチェーンのセキュリティは，米国連邦情報システムを対象としたセキュリティ管理策の基準である NIST SP800-53[3](民間企業向けには SP800-171[4])で触れられており，よりサプライチェーンに特化した SP800-161[5]も策定されている．一方で，サプライチェーン上のサプライヤー企業や製品，仕向け地などのセキュリティ水準の評価は，過去の取引実績や知名度などの主観に頼ることが多く，セキュリティ水準の低いサプライヤーが選択される可能性は否定できない．そこで，筆者らは複数の情報源から収集したオープンデータを活用し，セキュリティ水準を数値化するサプライヤー企業のセキュリティ評価手法を提案した[9]．この方式で，サプライチェーン上のサプライヤー評価方法の枠組みを示し，実装には詳細に言及していない．

本稿では，評価対象のサプライヤー企業に関するセキュリティニュースからクラスタリング分析により抽出したキーワード群を Social Networking Service(以下，SNS)や Web 検索エンジンの表示順序を利用して評価し，選定されたキ

1 三菱電機株式会社 情報技術総合研究所
Mitsubishi Electric Corporation,Information Technology R & D Center
* Mori.Takumi@db.MitsubishiElectric.co.jp

ーワード群を説明変数とした重回帰分析によりセキュリティ評価値を算出する方法を提案する．さらに，機械学習のアプローチを用いて実装の観点で評価を実施し，その実現可能性を示す．

2. 関連研究

サプライチェーンにおける情報セキュリティに関する取り組みとして，久保ら[6]は，日本企業を対象に，サプライチェーンのリスク管理の PDCA サイクルを提案している．その中で，サプライヤーのモニタリングには立ち入り監査が一般的であるが，コストの面から国際標準による認証が有効と言及している．しかし，それだけでは，サプライヤーによる情報セキュリティ事故の抑制は限定的であると述べられており，効果的なサプライヤー評価方式の必要性が示唆されている．原田ら[7]は，サプライチェーン全体の IT ガバナンスを調整し，統合する方法を提案している．この方式でも，サプライヤーのモニタリングと評価に言及しているが，具体的な方式は示されていない．そこで我々は，国際標準や業界標準を参考に作成された，青地らのサプライヤーのリスクマネジメントのフレームワーク[8]の一部のステップを技術的に解決する方法として，サプライヤーの情報セキュリティに関するオープンデータを情報源ごとに重みづけしてリポジトリ化し，セキュリティ評価要求に応じてセキュリティ評価値を算出する方式を検討した[9](図 1)．検討の結果，セキュリティ評価にオープンデータを使用することで説明可能なセキュリティ評価値を算出する枠組みは確立したが，具体的な情報源や実装への言及をしていない．

一方で，サプライヤーのセキュリティ評価は，一般的に企業が独自に行うことが多い．そこで効果的なサプライヤーのセキュリティ評価手法を検討するにあたり，取引先企業に対する属性確認や口座開設時のセキュリティチェック，CSR 評価に関する評価方法や判断基準を調査した．

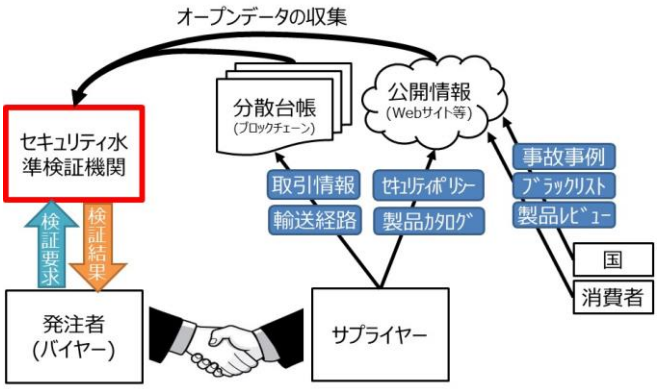

図 1 筆者らのサプライヤー企業の評価方法[9]

その結果，属性確認は東京商工リサーチの企業評点[a]や，日経テレコンの企業情報[b]，自社内のブラックリストなどの情報源から総合的に判断していることが分かった．また，評価に有料の情報や非開示の情報を参照する場合は，評価結果の検証(説明)が難しい．そこで我々は，「セキュリティ評価値が未知のサプライヤーのセキュリティ評価値を説明可能な形で算出すること」を課題に設定する．

3. 提案方式

本稿では，セキュリティ評価値を機械学習のアプローチで推定する方法を提案する．セキュリティ評価値が未知のサプライヤーの評価値を求める方法として有効な手法が，機械学習である．機械学習を利用したもっとも単純な方法として，取引可否が既知のサプライヤーをニューラルネットで学習して分類機を構築したうえで，未知のサプライヤーを判断する方法が考えられる．しかし，分類の過程がブラックボックスになるため，「説明可能な形で算出できる」という課題を解決できていない．そこで我々は，セキュリティニュースに含まれるキーワードのセキュリティ評価値に対する寄与率をもとに，セキュリティ評価値を検証(説明)するアプローチを採る．このアプローチを実現するための単純な方法は，重回帰分析を用いて，サプライヤー企業が信頼できるか否かを，複数の説明変数によって説明することである．

そこで，セキュリティニュースを重回帰分析し，サプライヤーのセキュリティを評価する方式(以降，本方式)を提案する(図 2)．始めに，セキュリティキーワードを特定し，ニュースサイトからニュースを収集する．次に，収集したニュースに含まれるキーワードを分析し，キーワード群(説明変数候補)を作成する．そして，Web 検索エンジンの表示順序や SNS による反響を考慮して説明変数候補からセキュリティ評価値を算出する重回帰式の説明変数を選択する．

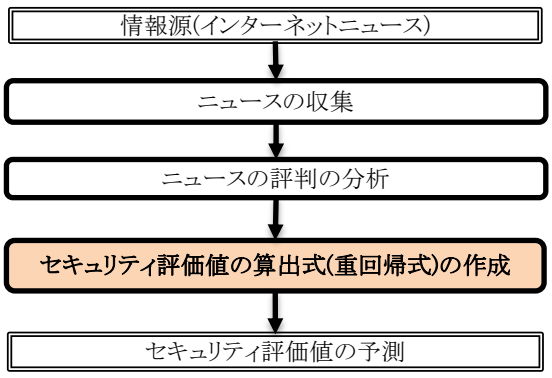

図 2 提案方式の流れ

a https://www.tsr-net.co.jp/service/product/national/evaluation/

b http://telecom.nikkei.co.jp/guide/menu/company/

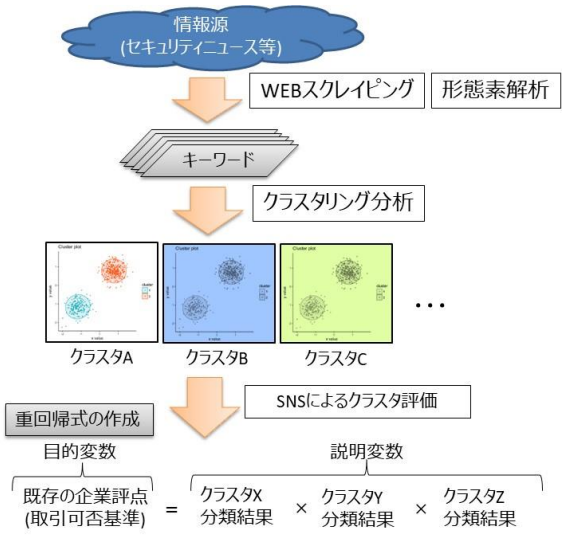

図 3 本方式の実装イメージ

重回帰式は，既存のセキュリティ評価(企業評点など)に回帰するように作成する．最後に，評価対象のサプライヤーに関するニュースを収集し，重回帰式にあてはめることで，セキュリティ評価値を予測する．このようにすることで，算出したセキュリティ評価値を反響の高いキーワードによって説明可能となり，課題を解決することができる．

図 3 は本方式の実装イメージである．以降では，本方式の説明と併せて実装のポイントも説明する．

3.1 ニュースの収集

セキュリティ評価に使用するニュースを収集するために，以下の方法で使用するキーワードを選定する．

- ***セキュリティ関連ニュースサイトに掲載されているニュースを形態素解析し，頻出キーワードを抽出***
- ***セキュリティ関連ニュースサイトに掲載されている注目キーワードを使用***

このように抽出したキーワードを用いて，学習対象の企業に関するセキュリティ関連ニュースを検索する．具体的には，既存の企業評点が存在する企業を選出し，「[企業名]＋[セキュリティキーワード]」でニュースを検索する．検索したニュースに共通的に含まれるキーワードを説明変数の候補とする．

この手順を実装する場合，まず，Web に公開されているセキュリティ関連ニュースを Web スクレイピング[10][11][12][13]の技術を用いて収集する．収集したニュース内容を形態素解析[14][15]によってキーワードを抽出する．次に，抽出したキーワード群を分析し，重要と判断したキーワード群を説明変数の候補とする．キーワード群の分析には，クラスタリング分析を用いる．ニュースを形態素解析して抽出したキーワード群を，クラスタリング分析して得られたクラスタ(以降，キーワードクラスタ)を説明変数の候補とする．

キーワードを分析する方法としてクラスタリング分析を採用した理由を詳述する．分析方法は大きくトピック分析とクラスタリング分析の 2 種類がある．トピック分析は，多くの文章情報から潜在的な意味(トピック)を把握するための手法である． LSA(Latent Semantic Analysis)[16]，PLSA(Propabilistic Latent Semantic Analysis)[17]，LDA (Latent Dirichlet Allocation)[18]が代表的な手法として存在する．これらの方法に共通するのは，文章中に潜在的にトピックが複数存在するということである．今回は，セキュリティ評価向けに情報源を分析するため，潜在的なトピックは「情報セキュリティ」に関連したものに限られる．そのため，期待した数のトピック(説明変数の候補)を得られるとは考えにくい．そのため，トピック分析は利用せず，クラスタリング分析を用いる．

クラスタリング分析を実施するにあたり，文章中に含まれる単語をベクトル化する．ベクトル化する手法には，TF-IDF(Term Frequency-Inverse Document Frequency)法[19][20]，Word2Vec[21]，Doc2Vec[22][23]，Fasttext[24]といった手法があるが，今回用いる情報源の特徴や評価のしやすさを考慮し，Word2Vec を単語のベクトル化方式として採用する．

ベクトル化した情報をクラスタリング分析する方式は，階層型クラスタリングと非階層型クラスタリングに大別される．階層型クラスタリングは，ビッグデータのようにクラスタリング対象が多い場合に計算量が爆発する欠点がある[25]．本方式はセキュリティニュースのようなビッグデータを扱うため，ビッグデータの分析に強い非階層型クラスタリングを採用する．非階層型クラスタリングには， k-means 法， DBSCAN (Density-based spatial clustering of applications with noise)[26]， Mean-shift[27]がある．DBSCAN や Mean-shift は予めデータの特性が分からない場合に不向きであるため，k-means を用いる．

3.2 ニュースの評判の分析

収集したニュースに対し，ニュースヘッドラインの Web ページ表示順序，ヒット数，及び SNS による「ツイート数」「リツイート(RT)数」「いいね数」等を調査し，以下の例のように反響の多いニュースを特定する．

- ***検索サイトヒット数 100 件以下は除外***
- ***ツイート数，RT 数，いいね数がすべて 0 のものは除外***

実装の観点では，検索エンジンに関する情報は Web スクレイピング，SNS に関する情報は Twitter API[28]などを使用する．ツイート，リツイート(RT)，いいね，が最も単純な指標だが，投稿内容を端的に表すキーワードであるハッシュタグや，ツイートがユーザに表示された回数を表す指標であるインプレッション，エンゲージメント(クリック，リツイート，返信，フォロー，いいね)の数をインプレッションの合計数で割って算出した指標であるエンゲージメント率を利用することで，より高精度な反響情報を得ること

ができる．

SNS から得られる情報から 3.1 で作成したキーワードクラスタを選定する方法は，以下のように行う．

手順1　キーワードクラスタの各クラスタに含まれるキーワードからいくつかを選択[c]

手順2　選択したキーワードを用いて Twitter API で検索し，ツイート，リツイート(RT)，いいね等の情報を収集

手順3　収集した情報を分析し，インプレッション，エンゲージメント率を計算し，これらの指標で反響が高いと判断されたキーワードクラスタを説明変数の候補として採用

3.3 セキュリティ評価値を求める重回帰式の作成

既存のセキュリティ評価値(企業評点や，自社の取引実績など)を目的変数，3.2 の処理によって選択されたクラスタを説明変数とする．セキュリティ評価値yの重回帰式を数式(1)に表す．但し，x_nは，キーワード群nのニュース数であり，β_mは偏回帰係数で，最小二乗法で求めたものである．

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (1)$$

例えば，東京商工リサーチの企業評点を目的変数，セキュリティニュースに含まれる共通的なキーワード群の出現数を説明変数とすると表 1 のようになる．また，表 1 から重回帰式を作成すると数式(2)のようになる．

$$y = 69.6319 + (-0.6283x_1) + 0.0042x_2 + (-0.0016x_3) \quad (2)$$

重回帰式は，目的変数にあたる情報が複数ある場合は複数作成する．重回帰式の作成が，本方式の性能に大きく影響する．そのため，別途 4 章にて評価を実施した．

3.4 評価対象のセキュリティ評価値の算出

未知のサプライヤー企業について，説明変数として用いたx_1～x_3のキーワードを含むニュースをカウントし，導出した複数の重回帰式に当てはめることで得られるセキュリティ評価値の平均値を最終的なセキュリティ評価値とする．

表 1 目的変数と説明変数の例

企業名	企業評点(目的変数)(y)	キーワード群(説明変数)		
		“脆弱性”，“JVN”,“CVE”(x_1)	“コマンド”(x_2)	“セキュリティ”(x_3)
A 社	69	0	5	392
B 社	63	4	372	5,180
C 社	69	0	125	3,490
D 社	53	2	170	3,480
E 社	63	1	185	4,690

例えば，サプライヤーA のセキュリティ評価値を計算する場合，表 2 に示したようにキーワード群を含むニュースが見つかったとする．これを数式(1)にあてはめると，サプライヤーA のセキュリティ評価値は 64.6533 点となる．

4. 評価

本章では，本方式が効果的に動作するパラメータを特定するための評価を行う．本方式は，機械学習の一般的な手順を踏襲しているため，期待する効果を得るには適切なパラメータ設定が不可欠である．特に重回帰式の作成がセキュリティ評価値の精度に大きく影響するため，重回帰式の作成に焦点を置き，評価を実施した．

4.1 評価環境

評価環境を表 3 に示す．ただし，評価に必要な主たるものとし，依存ライブラリは記載していない．

評価データは，特定のサプライヤー企業に関する情報を利用することが望ましかったが，サプライヤー企業に関する情報(セキュリティニュース)が現時点では十分に公開されていない．そこで，ある製品の仕向け地を対象とした取引国の分類に関する重回帰式の作成するケースを想定し，特定の国に関するセキュリティニュースを評価データとしてセキュリティ評価値の算出を試行した．

表 2 未知のサプライヤーA に関する情報

企業名	キーワード群(説明変数)		
	“脆弱性”，“JVN”,“CVE”(x_1)	“コマンド”(x_2)	“セキュリティ”(x_3)
サプライヤーA	3	100	2150

表 3 評価環境

種別	S/W	備考
プラットフォーム	Anaconda 3.6 64bit(Anaconda3-2019.10)	開発ツール Jupyter Notebook 6.0.1，数値計算用に numpy 1.16.5，機械学習用に scikit-learn 0.21.3 を含む
開発言語	python 3.7.4	-
スクレイピング	requests 2.22.0 selenium 3.141.0 beautifulsoup4 4.8.0	selenium の WebDriver には Firefox に対応する geckodriver 0.26.0 を使用
形態素解析	mecab 0.996 (64bit) mecab-python-windows 0.996.3 collections 3.6	64bit 版 MeCab は有志によるビルド[d]を使用 collections 3.6 は辞書用
自然言語分析	genism 3.8.1	LDA 及び Word2Vec を使用する(詳細は後述)

c 簡易的な実験の結果，キーワードは 3 程度までであれば，反響情報を得やすいことがわかっている

d https://github.com/ikegami-yukino/mecab/releases/tag/v0.996

4.1 重回帰式のチューニング

ある製品の仕向け地を評価対象とし,「国」レベルで仕向け地を評価する．経産省の輸出管理上における国別カテゴリー[e]と制裁情報[f]を参考に禁輸国，注意国，ホワイト国を定義し，これを目的変数とした重回帰式を作成する．なお，Twitter API によるキーワードの反響調査は分析に時間を要することから，机上検討のみとした．評価対象の仕向け地は以下のとおりである．

禁輸国　：　ウクライナ，イラク，スーダン
注意国　：　ロシア，アラブ，イスラエル，中国
ホワイト国　：　チェコ，オーストラリア，英国，ドイツ

仕向け地の評価を目的変数，ニュースのキーワードクラスタを説明変数とした重回帰式を以下の手順で作成した．

手順1　セキュリティニュースの収集
(1) ScanNetSecuritygの最新ニュース 1050 件(2020/3/3 時点)を収集(ニュースクラスタ用)
(2) ScanNetSecurity で各国名をキーワードに検索し，上位 10 件のニュースを収集(仕向け地クラスタ用)
手順2　収集したニュースのクラスタリング[h]
(1) ニュースクラスタ用データからクラスタを作成
(2) 仕向け地クラスタ用データからクラスタを作成
手順3　ニュースクラスタへの仕向け地のマッピング
(1) ニュースクラスタに対応するキーワードと，仕向け地クラスタに対応するキーワードを比較し，マッピングを実施
(2) 仕向け地ごとのクラスタに含まれるキーワード数の特性を考慮し，説明変数用のクラスタを選定
手順4　選定した説明変数から重回帰式を作成

表 4 に手順 3 の実施結果を示す. 仕向け地の特性(禁輸国，注意国，ホワイト国)ごとに分布に差があれば，回帰式の説明変数として採用できる．

表 4 ニュースクラスタへの仕向け地のマッピング

		目的変数	説明変数候補									
			C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
禁輸国	ウクライナ	2	2	4	0	0	0	1	3	1	2	0
	イラク	2	1	5	0	1	1	2	1	0	4	0
	スーダン	2	3	2	0	1	0	0	0	0	0	0
注意国	ロシア	1	2	5	1	0	0	1	1	0	0	1
	アラブ	1	2	4	0	0	0	1	3	1	2	0
	イスラエル	1	1	4	2	1	0	2	3	0	3	0
	中国	1	5	35	7	8	5	6	7	0	7	1
ホワイト国	チェコ	0	2	2	2	1	0	3	0	0	2	0
	オーストラリア	0	0	4	0	1	1	1	0	1	0	0
	英国	0	0	2	1	2	1	2	0	0	1	0
	ドイツ	0	1	6	0	2	1	2	2	0	2	0

目的変数：禁輸国=2, 注意国=1, ホワイト国=0

そこで，クラスタの分布を仕向け地の特性ごとにレーダーチャートで可視化したものを図 4 に示す(外れ値の中国は除外)．禁輸国，注意国，ホワイト国の各チャートの形が異なる場合，説明変数(クラスタ)が独立していると考えられるが，チャートに特徴が表れなかった．また，クラスタすべてを説明変数として重回帰式を作成しても，多重共線性により，説明変数の P-値が計算できなかった．

さらに仕向け地の特性と各クラスタに対応するキーワード数を精査し，目視により特徴が出たクラスタ 1,3,4,5,6,7,9 に絞った結果が図 5 である．仕向け地の特性によりレーダーチャートの形が異なっており，これらのクラスタが説明変数の候補として有効と判断した．

次に，クラスタ 1,3,4,5,6,7,9 を説明変数とした場合の重回帰式を数式(3)に，その評価結果を図 6 に示す．但し，C_nKw は，クラスタ n のキーワード数である．

$$\begin{aligned}\text{目的変数} &= 1.3699 + 0.3094\times C_1Kw \\ &\quad + 0.3411\times C_3Kw + (-0.2674)\times C_4Kw \\ &\quad + 0.3805\times C_5Kw + (-1.2505)\times C_6Kw \\ &\quad + (-0.2312)\times C_7Kw + (0.7228)\times C_9Kw\end{aligned} \tag{3}$$

今回作成した重回帰式(数式(3))の決定係数は図 6 中の「重決定 R2」から 0.90 であることがわかる．これは，説明変数が目的変数を 90%説明していることを表す．決定係数は，1 に近づくほど目的変数の選択がもっともらしいと判断できるため重要な指標となる．しかし，決定係数の良し悪しに統計学的基準はないため，重回帰式を比較評価する時の指標とする．P-値は説明変数の信頼度を表し，統計的仮説検定により説明変数による値の算出が偶然によるものか否かを判定している．数式(3)は，クラスタ 1,6,9 の P-値が小さく，今回のデータでは，目的変数との関連性が大きいことがわかる．

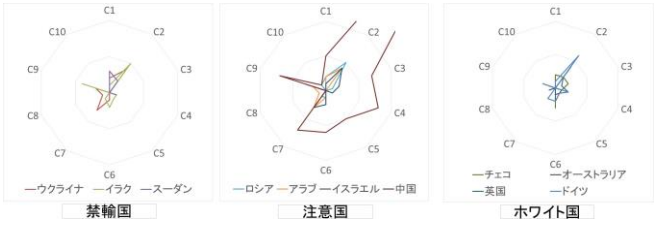

図 4 クラスタ分布(全クラスタ)

図 5 クラスタ分布(選定後)

e https://www.meti.go.jp/press/2019/08/20190802001/20190802001.html
f https://www.meti.go.jp/policy/external_economy/trade_control/01_seido/04_seisai/seisai_top.html
g https://scan.netsecurity.ne.jp/
h クラスタ数はデータセットから決定すべきだが，今回は 10 で固定.

回帰統計	
重相関 R	0.952261
重決定 R2	0.906802
補正 R2	0.68934
標準誤差	0.46329
観測数	11

分散分析表

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	1.369912	0.484046	2.830127	0.066183	−0.17053906	2.910364	−0.17054	2.910364
クラスタ1	0.309359	0.176172	1.756009	0.177347	−0.25129779	0.870016	−0.2513	0.870016
クラスタ3	0.341123	0.274952	1.240662	0.302913	−0.53389773	1.216143	−0.5339	1.216143
クラスタ4	−0.26741	0.359565	−0.7437	0.51101	−1.41170416	0.87689	−1.4117	0.87689
クラスタ5	0.380529	0.456788	0.833054	0.465934	−1.07317475	1.834233	−1.07317	1.834233
クラスタ6	−1.25048	0.410834	−3.04376	0.055702	−2.55793385	0.056978	−2.55793	0.056978
クラスタ7	−0.23124	0.151925	−1.52205	0.225357	−0.71473096	0.252256	−0.71473	0.252256
クラスタ9	0.722804	0.209602	3.448463	0.040982	0.055757647	1.389851	0.055758	1.389851

図 6 重回帰式の評価

参考までに，今回得られた回帰式の評価(図 6)から，P-値の低いクラスタのみを抽出して，再度重回帰式を作成したところ，決定係数が低下した．P-値が低く，決定係数を高くする説明変数の選択をすべきであるが，本評価環境では明確な傾向を把握できなかった．

5. 考察

評価結果から，本方式の実現可能性について考察する．まず，即時性，多様性，検索性が高いセキュリティ関連のインターネットニュースを情報源として採用するうえでは，同一トピックに対する情報量の多さとコンテンツの裏付けの不十分さをカバーする必要がある．そこで，インターネット検索エンジンや，特定の情報に対する個人の意見を収集しやすい SNS を使い，セキュリティ評価値を算出する重回帰式の精度を高めることができる可能性があることがわかった．ニュースのようなビッグデータには，機械学習のアプローチが有効であり，既存の正解データ(企業評点など)へ回帰させることで，セキュリティ評価値を算出できる．本方式で正確なセキュリティ評価値が予測できるかは重回帰式の精度に依存する．検討の結果，より多くの情報を含む長い文章を持つ偏りのないセキュリティニュースを情報源とし，クラスタリング分析による説明変数候補の抽出，SNS による説明変数の選択によって品質の高い重回帰式が得られる可能性が高いことがわかった．

本評価環境では，高い決定係数と低い P-値を含む説明変数から構成される重回帰式が得られているため，セキュリティニュースが情報源として適切であるといえる．しかし，重回帰式は説明変数の選択によって決定係数が大きく変化するため，説明変数の選択には多くの試行錯誤が必要となる．これを手動で行う場合，評価者の主観や世情が大きく反映される恐れがあるが，SNS で反響の大きいキーワードを含むクラスタから説明変数を選択することで，主観の混入や試行錯誤の削減が可能になる．SNS による説明変数の選択については，各クラスタの上位キーワード 3 つ程度を SNS で検索することで，キーワードの反響情報を得られることが追加調査の結果からわかっているが，本方式への組み込みについては追加の評価が必要である．

6. まとめ

本稿では，サプライチェーンにおけるサプライヤーのセキュリティ評価を自動で実施するための，Web データの回帰分析によるセキュリティ評価手法を提案した．また，セキュリティ評価値を検証可能な形で算出するという課題に対し，本評価環境下では，本方式が有効かつ実装が可能であることを示した．さらに，本方式におけるセキュリティ評価値の算出精度を向上させるパラメータ調整方針を明らかにした．本方式を実用レベルまで高めるには，より高精度のチューニングが必要である．そのためには，近年研究が盛んに行われている説明可能な AI(XAI:Explainable Artificial Intelligence)[29]を用いることも有効と考えられる．

参考文献

[1] KASPWRSKY:2018 年サイバー脅威の予測, KASPERSKY (online), available from <https://media.kaspersky.com/jp/pdf/pr/Kaspersky_KSB2017_Predictions-PR-1043.pdf> (accessed 2019-3-28).

[2] KASPWRSKY:2020 年 高度なサイバー脅威の予測, KASPERSKY(online), available from <https://media.kaspersky.com/jp/pdf/pr/Kaspersky_KSB2019_Predictions-PR-1052.pdf> (accessed 2020-4-15).

[3] NIST: SP800-53 Rev.4 Security and Privacy Controls for Federal Information Systems and Organizations, NIST(online), available from <https://csrc.nist.gov/publications/detail/sp/800-53/rev-4/final>(accessed 2019-3-28).

[4] NIST: SP800-171 Rev.1 Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations, NIST(online), available from <https://csrc.nist.gov/publications/detail/sp/800-171/rev-1/final>(accessed 2019-3-28).

[5] NIST: SP800-161 Supply Chain Risk Management Practices for Federal Information Systems and Organizations, NIST(online), available from <https://csrc.nist.gov/publications/detail/sp/800-161/final >(accessed 2019-3-28).

[6] 久保知裕, 原田要之助: サプライチェーンにおける情報セキュリティの研究,情報処理学会研究報告, Vol.2014-EIP-65, No.3, pp.1-8(2014).

[7] 原田要之助, 久保知裕: 複数企業にまたがった IT サービスのサプライチェーンにおける IT ガバナンスの課題について, 情報処理学会研究報告, Vol.2015-EIP-67, No.3, 1-8(2015).

[8] 青地忠浩: サプライチェーンリスクマネジメントのフレームワークと実例, 日本 LCA 学会誌, Vol.14, No.4, pp.256-266(2018).

[9] 森拓海, 藤田真浩, 山中忠和: オープンデータを用いたサプライヤーのセキュリティ評価手法, コンピュータセキュリティシンポジウム 2019 論文集 (2019), pp.92-97(2019)

[10] Gray, M.:Internet Growth and Statistics: Credits and Background, MIT(online), available from <www.mit.edu/people/mkgray/net/background.html>(accessed 2020-4-28).

[11] Richardson, L.:Beautiful Soup, Crummy(online), available from <https://www.crummy.com/software/BeautifulSoup/>(accessed 2020-5-13).

[12] Reitz, K.:Requests: HTTP for Humans, A Kenneth Reitz Project(online), available from <https://requests.readthedocs.io/en/master/>(accessed 2020-5-13).

[13] Huggins, J.: SeleniumHQ Browser Automation, Software Freedom Conservancy(online), available from <https://www.selenium.dev/>

(accessed 2020-5-13).
[14] 工藤拓: MeCab: Yet Another Part-of-Speech and Morphological Analyzer, github(online), available from <https://taku910.github.io/mecab/>(accessed 2020-5-13).
[15] 打田智子: Janome v0.3 documentation (ja), github(online), available from <https://mocobeta.github.io/janome/>(accessed 2020-5-13).
[16] Deerwester, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R.:Indexing by Latent Semantic Analysis, Journal of the American Society for Information Science, 41(6), 321-407(1990).
[17] Hofmann, T.: Probabilistic Latent Semantic Analysis, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp.50-57(1999).
[18] Blei, D.M., Ng, A.Y., and Jordan, M. I.: Latent Dirichlet allocation, Journal of Machine Learning Research, vol.3, pp.993-1022(2003).
[19] Robertson, S.E., Sparck-Jones, K.:Relevance weighting of search terms,Journal of the American Society of Information Science, 27, pp.129-146(1976).
[20] Salton, G., Buckley, C.:Weighting approaches in automatic text retrieval, Information Processing and Management, 24(5), pp.513-523(1988).
[21] Mikolov, T.,Chen, K., Corrado, G., Dean, J.:Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781(2013).
[22] Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents, Proceedings of The 31st International Conference on Machine Learning (ICML2014), pp.1188-1196(2014).
[23] Řehůřek, R.:gensim Doc2Vec Model, gensim(online), available from <https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html#sphx-glr-auto-examples-tutorials-run-doc2vec-lee-py>(accessed 2020-5-14).
[24] Facebook:FastText, Facebook Research(online), available from <https://research.fb.com/downloads/fasttext/>(accessed 2020-5-14).
[25] 神蔦敏弘:データマイニング分野のクラスタリング手法(1) : クラスタリングを使ってみよう!, 人工知能学会誌, Vol.18, No.1, pp.59-65(2003).
[26] Ester, M.,Kriegel, H.P., Sander, J., Xu, X.:A density-based algorithm for discovering clusters in large spatial databases with noise, KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp.226-231(1996).
[27] Fukunaga, K., Hostetler, L.D.:The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition, IEEE Transactions on Information Theory, 21(1), pp.32-40(1975).
[28] Twitter:Docs-Twitter Developers, Twitter(online), available from < https://developer.twitter.com/ja/docs>(accessed 2020-5-14).
[29] Adadi,A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access (Volume:6), pp.52138-52160(2018).