明治大学総合数理学部

**2018** 年度

卒　業　研　究

# 企業プレスリリースからのサイバーインシデント情報の自動収集と分析

学位請求者　先端メディアサイエンス学科

池上和輝

# 目次

# 第 1 章

# はじめに

## 1.1　研究背景

近年，企業における内部不正や外部からの攻撃による個人情報漏洩などのサイバーインシデントが増加している．インシデントによる，データの損失や破損，事業妨害，賠償責任など，企業は一つのインシデントに対して大きな損害を被る可能性がある．企業の経営者たちは個人情報を扱う上でセキュリティのマネジメントが必要とされている．これらの損害をカバーするマネジメントの一環に，保険会社が提案しているセキュリティ保険がある．しかし，保険料を決めるにあたっての正確な被害額算出モデルが確立されていない．被害額を算出するモデルには，日本ネットワークセキュリティ協会 JNSA の JO モデル [1] や明治大学大学院の山田らが提案したモデルなどがある．これらのモデルの問題点の一つに，モデル作成の上で用いるインシデントが網羅的に収集できていないという点がある．

## 1.2　研究目的

本研究では，まずインシデントの傾向と特徴を明らかにするために独自のインシデントデータセットを作成し漏洩インシデントの傾向や特徴を明らかにすること，サイバーインシデントをメディアなどによる偏りなく網羅的に自動で収集して分類することを目的とすることの二点を研究目的とする．

## 1.3　本論文の構成

本論文では二章で，インシデントの傾向と特徴を明らかにするために手動によるインシデント調査の結果，三章ではそれらの結果を踏まえたクローラー・自動分類システムの開発について示す．四章でそれらについての考察を述べる．

# 第 2 章

# 手動によるインシデント調査

## 2.1　データセットの作成・分析

### 2.1.1　データセットの作成

朝日新聞の記事検索システム「聞蔵 Ⅱ」[4] を用いて 2015 年の漏洩インシデントの情報を収集した．「情報漏えい＋紛失＋漏洩＋不正アクセス＋誤送信＋盗難」の検索語を用いた．我々の調査では，JNSA では不足している，社内規則違反の有無や，報道時点での流出の可能性の有無，情報漏洩が起きてから報道されるまでの期間などの新しい要素を加えている．

表 2.1 に収集した漏洩インシデント数と JNSA を比較する．企業名と漏洩件数，公表日について両者が一致しているかを判断した．比較の結果の約半分が JNSA には含まれない新規のデータであった．逆に JNSA の件数が多いのは，JNSA のデータセットでは企業の支店まで分けてデータセットに加えていることが JNSA のデータセットの件数が多い要因の一つだと考えられる. また，JNSA ではインターネット上に公開されたインシデントを対象として収集していることもインシデント数の違いの原因だと考えられる．

表 4.1 に本研究と JNSA データセットの例を示す．項目の*印は本研究のみの要素を示している．JNSA のデータッセットには，持ち会社/親会社/関連会社の情報，管理委託先情報の他に，社会的責任度/インシデント内容要約/事後対応姿勢/経済的ランク/精神的ランク/基礎点/責任/対応/特定/一人当たりの損害賠償額が含まれる．社会的責任度以降の値は，JO モデルによる損害賠償額を求めるための値であり，表 4.1 の最終行以外の情報から算出する．

インシデントの公表日は新聞の発行日ではなく，記事中で述べられる公表日と漏洩日である．社内規則違反と二次被害については可能性が否定されていない事件を「有」と分類した．新聞の記事に「漏洩の可能性が極めて低い」,「誤って破棄した可能性が高い」,「一時的な紛失」などの記載がある場合は，二次被害の可能性が低いと判断する．

### 2.1.2　データセットの比較と分析

表 4.2 に収集した漏洩インシデントと JNSA の漏洩レコード件数の統計を示す．最大値はどちらも日本年金機構のものであるが，本調査では新聞記事から「基礎年金番号と氏名」が約 3 万 1 千件,「番号，氏名，生年月日」が約 116 万 7 千件のように各要素が何件漏洩したかという情報と，個人の漏洩した要素は不明だが 101 万人の個人情報が漏洩したという情報を得た．本調査では漏洩原因と漏洩要素の関係を明らかにすることも目的の一つなので，前者の情報をデータセットに加えた．そのため，最大値に差が出た．平均値の差は最大値の

表 2.1　データセットのインシデント数の比較 [件]

| JNSA | 本データセット | 共通 |
|---|---|---|
| 788 | 279 | 145 |

表 2.2　データセットの例

| 項目 | JNSA | 本調査 | 共通事例 |
|---|---|---|---|
| 公表日 | 2015/10/6 | 2015/9/7 | 2015/6/1 |
| 発生日* | 不明 | 2015/9/2 | 2015/5/8 |
| 企業名 | ゴールドボンド | 福岡県 | 日本年金機構 |
| 業種 | サービス業 | 公務 | 公務 |
| 持ち会社/親会社/関連企業 | 広島県 | 不明 | |
| 管理委託先 | 委託先あり | 不明 | |
| 漏洩件数 | 77 | 35 | 1160000 |
| 漏洩原因 (件) | 誤操作 | 紛失・置き忘れ | 不正アクセス |
| 漏洩経路 | 電子メール | 紙 | インターネット |
| 漏洩要素 | 氏名/電話番号 | 氏名/住所 | 氏名/ID |
| 社内規則違反* | 不明 | 0 | 1 |
| 被害の可能性* | 不明 | 0 | 1 |
| 報道までの日数* | 不明 | 5 日 | 24 日 |
| 社会的責任度など | 有 | 不明 | 有 |

表 2.3　漏洩した個人情報レコード数の統計量

| | 平均値 | 最大値 | 中央値 | 最頻値 |
|---|---|---|---|---|
| 本調査 | 8104 | 1160000 | 54 | 1 |
| JNSA | 6680 | 1014653 | 117 | 1 |

差が影響していると考えられる．最頻値がどちらも 1 の一方で，中央値が JNSA のほうが大きいことから規模の大きな事件をより収集していたことがわかる．最大値と平均値が高いが，中央値が 54 件で最小値が 1 件のことから，回数は少ないが規模の大きい事件が複数回あったことがわかる．

表 2.4 に，本調査のデータセットと JNSA データセットの漏洩原因ごとの件数を示す．また，その時の各要素の割合を図 2.1 と図 4.3 に示す．発生する件数の多い漏洩原因の上位 5 つは，順位が同じだった．しかし，図 2.1 と図 4.3 で比較すると本データセットでは紛失・置き忘れの比率が約 6 割を占めており，管理ミス，誤操作の割合は JNSA に比べて小さい割合となった．また，不正アクセスと盗難は各データセットで件数は異なるのものの，全体を占める割合はほとんど近い値となった．発生することの少ない漏洩原因の設定ミス，ワーム・ウィルス，内部犯罪・内部不正行為，バグ・セキュリティホールなどは，本調査の収集では 1,2 件もしくは収集できなかった．これらのことから，1 種類の新聞記事から情報を収集すると盗難や不正アクセスなど事件性の高いものは，取り上げられやすいため全体の割合に変化がないが，それ以外の要素には偏りがでる，もしくは収集できないことがある．

表 2.4　漏洩原因の比較（件）

| 漏洩原因 | JNSA データセット | 本データセット |
|---|---|---|
| 管理ミス | 144 | 28 |
| 誤操作 | 196 | 39 |
| 紛失・置忘れ | 243 | 163 |
| 盗難 | 44 | 16 |
| 不正な情報持ち出し | 37 | 8 |
| 不正アクセス | 64 | 21 |
| 設定ミス | 15 | 2 |
| ワーム・ウィルス | 10 | 1 |
| 内部犯罪・内部不正行為 | 17 | 0 |
| バグ・セキュリティホール | 12 | 0 |
| その他 | 6 | 1 |

## 2.2　漏洩原因と漏洩要素の連関規則

### 2.2.1　漏洩原因と漏洩要素の連関規則

表 2.5 に R パッケージ”arules”により抽出した漏洩原因と漏洩要素の連関規則の一部および，各漏洩原因の件数，その代表的な事例を示す．規則には，本調査で収集したデータのみを用いた．「属性 A を持つ事例は属性 B を持つ傾向にある」という知識を連関規則という．lhs は条件，rhs は結論である．support は lhs と rhs の同時確率 P( lhs , rhs )，すなわちインシデント全体で lhs と rhs が同時に起きる確率を表す．confidence は条件付確率 P(rhs ｜ lhs )，lhs が起きたときに rhs が起きる確率を表している． また，lift は改善率といい，confidence が rhs の起こる確率の何倍かを示す．lift が 1 を超えない規則は有用でない [5].

表 2.5 では，lhs に漏洩原因，rhs に漏洩要素を制約させることで，原因と要素から成るルールについて調べる．規則 No.1 の support は全インシデントのなかで,「紛失・置き忘れによる氏名の漏洩が 55 %の確率で起きる」ことを示す．また，confidence は「紛失・置き忘れが原因で漏洩が置きた時, 95 %の確率で氏名が漏洩する」ことを意味する.No.3 の規則から「誤操作による漏洩があった時，約 60 %の確率でメールアドレスが漏洩する」ということがわかる．

次に，表 2.6 に表 2.5 と同様にして抽出した漏洩要素同士の連関規則の一部を示す．No.3 は氏名と住所が漏洩した時に，電話番号が漏洩するという規則である．No.4 から，電話番号と氏名が全インシデントの中で同時に漏洩する確率は低いが，電話番号が漏洩したときに氏名も漏洩している確率は約 97 %ということがわかる．つまり，電話番号の漏洩には氏名漏洩のリスクが高いことがわかる．同様のことが，No.2,4,5 の規則からも言える．

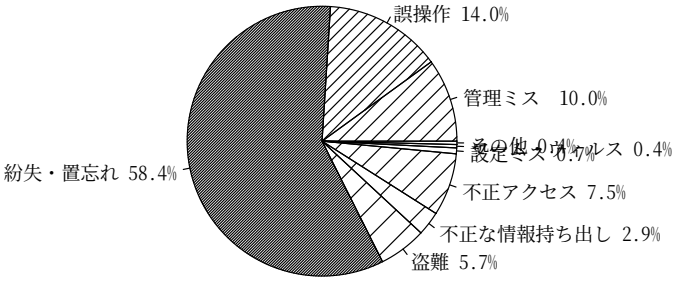


図 2.1　本調査

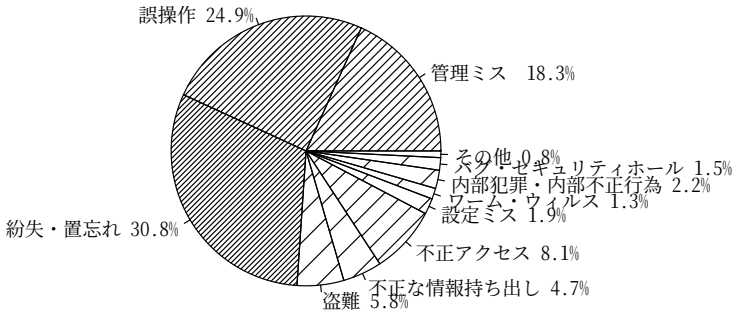


図 2.2　JNSA

## 2.3　事後対応の分析

### 2.3.1　報道までの日数

図 2.3 に報道までの日数のヒストグラムを示す．間隔は 1 10 になっている．0 日から 10 日かかるものが最も多い．その後は約 5 分の 1，約 3 分の 1，約 2 分の 1 のようにインシデント数が減少している．100 日を超

表 2.5　漏洩原因と漏洩要素の連関規則

| No. | lhs | rhs | support | confidence | lift | 件数 | 例 |
|---|---|---|---|---|---|---|---|
| 1 | 紛失・置忘れ | 氏名 | 0.557 | 0.957 | 1.117 | 156 | タカラトミー |
| 2 | 紛失・置忘れ | 住所 | 0.289 | 0.497 | 1.104 | 81 | TKC |
| 3 | 誤操作 | メールアドレス | 0.079 | 0.595 | 4.757 | 22 | 愛媛県 |
| 4 | 管理ミス | 氏名 | 0.089 | 0.926 | 1.080 | 25 | 長崎大学病院 |
| 5 | 管理ミス | 住所 | 0.057 | 0.593 | 1.317 | 16 | 静岡ガス |
| 6 | 不正アクセス | クレジット情報 | 0.014 | 0.190 | 13.334 | 4 | 日本年金機構 |
| 7 | 不正アクセス | ID/パスワード | 0.014 | 0.190 | 7.619 | 4 | 新日本プロレスリング |

表 2.6　漏洩要素同士の連関規則

| No. | lhs | rhs | support | confidence | lift |
|---|---|---|---|---|---|
| 1 | 氏名 | 住所 | 0.446 | 0.521 | 1.157 |
| 2 | 住所 | 氏名 | 0.446 | 0.992 | 1.157 |
| 3 | 氏名/住所 | 電話番号 | 0.154 | 0.344 | 1.529 |
| 4 | 電話番号 | 氏名 | 0.218 | 0.968 | 1.130 |
| 5 | 口座情報 | 住所 | 0.032 | 0.968 | 1.123 |

えるインシデントも数件確認した．また，平均は 11 日になった．（また，報道までに 378 日かかった事件があったが，図をわかりやすくするために省略する）．

### 2.3.2　報道までの日数と漏洩規模

発生から報道までにかかる日数と漏洩した個人情レコード件数の関係を図 2.4 に示す．（報道までに 100 日以上かかった事件が 3 件あったが，図をわかりやすくするため省略する）．比較的規模の小さい漏洩ほど，報道までの時間がかかることがわかる．一方で，規模の大きい漏洩は会社の信用にも大きく関わるため，すぐに報道されると考えられる．また，平均日数は 11 日であったが報道までの日数の上位 3 件のみが 100 日を超えており，それらにより平均日数が少し長めになったことが考えられる．

図 2.5 に報道までの日数別，漏洩規模の割合を示す．大規模と小規模の基準は漏洩件数の平均を基準に判断した．報道までの日数が 11 日に対して，30 日までは大規模割合の増加が確認できる，このことから大規模な漏洩が平均日数以上に報道までの日数を要することがわかる．しかし，30 日以降は大規模な漏洩が 0 になり，50 日以降で割合が増加するものの $20 \leq day \leq 29$ よりも割合が小さい．したがって，大規模な漏洩は報道までに平均日数以上を要するが，ほとんどのインシデントが約 1 ヶ月以内には報道される傾向がわかる．また，表 fig4 と表 fig5 から，漏洩日数が 1 ヶ月かかるインシデントは少ないものの，その 9 割が小規模な漏洩である．

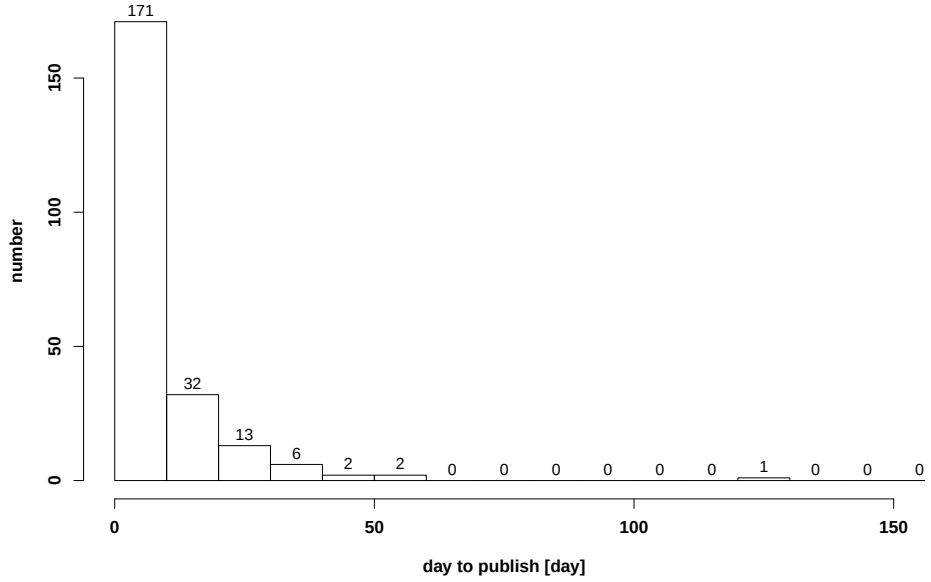


図 2.3　報道までの日数のヒストグラム

| | 31 日以上 | 30 日以下 |
|---|---|---|
| 全体 | 5 | 95 |

表 2.7　報道までに 1 ヶ月かかるインシデントの割合 [%]

| 小規模 | 大規模 |
|---|---|
| 90 | 10 |

表 2.8　31 日以上かかるインシデントの漏洩規模割合 [%]

## 2.4　考察

本章では，2015 年の新聞記事からデータッセットを作成し，漏洩原因についてデータマイニングを行った．一つのメディアから収集すると規模の小さなデータセットになり，一部の漏洩原因に偏りが出ることがわかった，一方で事件性の高い漏洩原因は全体の割合から見ると同じくらい収集することができた．また，新聞記事からは JNSA のデータセットになかった要素を 4 つ加えることで報道されるまで傾向を発見できた．規模の大きさや，漏洩原因が幅広く収集できている点では JNSA データセットのほうが優れていた．本調査のデータセットは JNSA にない要素から新しい傾向を発見できた点が優れていた．連関規則からは，漏洩原因が紛失・置忘れ，管理ミスのとき約 90 %の確率で氏名がが漏洩し，同様に誤操作の時は約 60 %の確率でメールアドレスが漏洩していることがわかった．また，要素間の連関規則では電話番号のように，「それ自体が漏洩する可能性は低いが，電話番号が漏洩した時には氏名が約 90 %の確率で同時に漏洩する」という要素があった．このことから，それらの要素は氏名と一緒に管理おり，漏洩による個人特定のリスクが高まる．漏洩規模と報道までの期間から，大規模な漏洩の報道は平均日数以上かかるものの 1 ヶ月以内には約 9 割が報道される傾向を発見した．また，30 日以上かかるインシデントは全体の約 5 %であったがそのうち約 90 %が小規模な漏洩

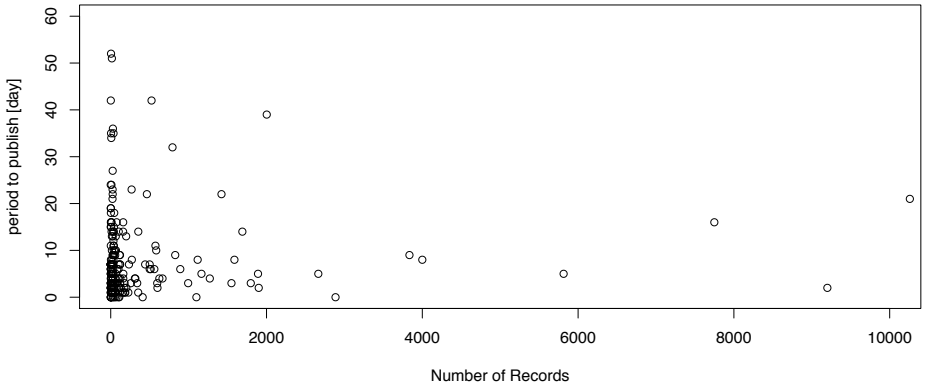


図 2.4　報道までの日数と漏洩の規模の散布図

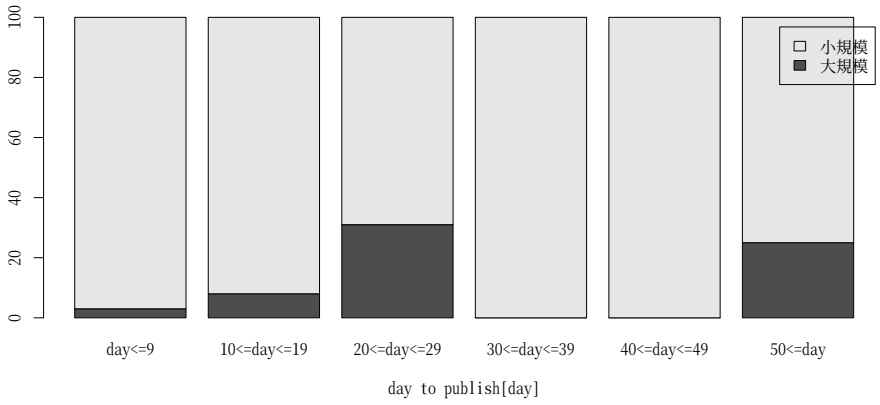


図 2.5　報道日数別漏洩規模の割合

だった．これらのことから，企業の事後対応として，大規模な漏洩では日数が平均よりかかるが 1 ヶ月以内に報道し，小さい報道では平均日数以内で報道するものがほとんどである．漏洩日数が 1 ヶ月以上かかる事後対応のインシデントは小規模であることが多い．

# 第 3 章

# クローラー・自動分類システムの開発

## 3.1　クローラー開発

システムの全体構成を図 3.1 に示す．1. 企業ウェブサイトの URL 与える．2. ウェブサイトの html とそのページ内のリンクを収集する．3. 収集したコンテンツをテキストに変換する．4. 特定のキーワードを含むテキストを保存する．

キーワードには，[6] の調査結果に基づき情報漏洩に関するプレスリリースに含まれるであろう単語，「お詫び」,「漏えい」,「漏洩」を使用した．2 ではリンク先ページについても任意の回数繰り返し，URL を収集した各企業について 1-4 を繰り返し行う．本調査では 2 を 3 回繰り返した．

自動分類システムの処理の例を図 3.2 に示す．日付，被害人数は正規表現により抽出する．漏洩原因は，次の手順で推定する．1. 特定の単語の出現頻度とその単語を含む文書の出現頻度から定めた tf-idf 値を用いて，各漏洩原因の文章の特徴語を抽出する．2. 1 で抽出した各漏洩原因の特徴語から成る 49 次元の特徴語ベクトルを作成する．3. 未知原因のプレスリリースを参照し，2 で求めた特徴語の TF 値を求める．4. プレスリリースから抽出した特徴語ベクトルと各漏洩原因ごとの特徴語ベクトルの cos 類似度を求める．最も高かった漏洩原因を出力する．表 3.1 に使用した特徴語と，各漏洩原因での出現頻度を示す．

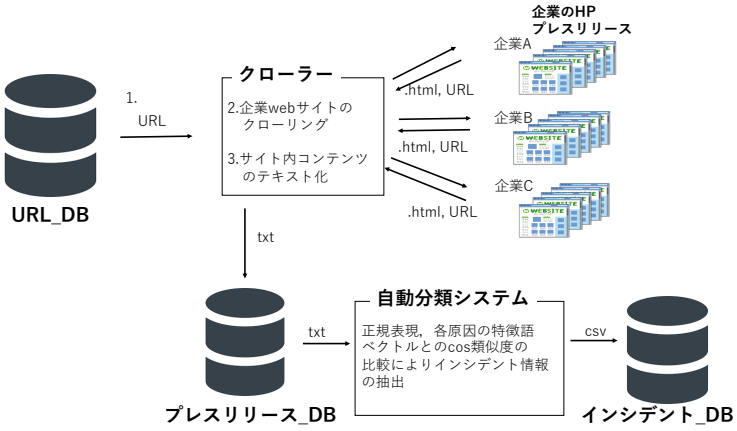


図 3.1　システム構成図

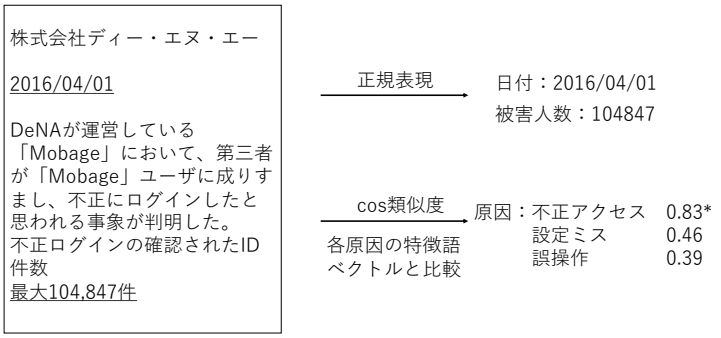


図 3.2　自動分類構造

## 3.2　調査対象企業

表 3.2 に調査対象企業の企業規模の内訳けと，セキュリティマネジメントの統計情報を業種別に示す．株式会社東洋経済新報社は，上場企業全社および主要未上場企業に調査票を送付し，その回答から社会的責任投資 CSR データベース [7] を作成している．本研究では各種セキュリティマネジメントの効果を検証するため，CSR データベース内の 537 企業を調査対象とした．537 企業のうち ISMS を導入しているのは 169 社 (31%)，CIO を設置しているのは 150 社 (28%) である．各企業の従業員数を元に，中小企業（従業員数 <300），大企業 1（従業員数 <1500），大企業 2（1500≤ 従業員数）の 3 種類に分類した．業種の中では情報通信・サービスその他が最も多く全体の約 20% である．

## 3.3　実験結果

表 3.3 に実験結果を示す．2018 年 11 月，2.2 節の企業ウェブサイトに本システムを用いてインシデント収集・分類を行なった．取得記事のうちインシデントに関する記事は 191 件であり，全体の 1% であった．また，191 件にはインシデント詳細ページとインシデント一覧ページの重複があったため，一意なインシデントは 179 件である．

表 3.4 にインシデントの日付，被害人数，漏洩原因の推定精度を以下の適合率で示す．

$$\text{適合率} = \frac{\text{正しく要素を抽出できたインシデント}}{\text{抽出した全インシデント}}$$

日付，人数などは全て 7 割を超えるが，全てを組み合わせた時の適合率は 50% であった．

表 3.1　各漏洩原因の特徴語ベクトル

| 特徴語 | 紛失・置忘れ | 管理ミス | 盗難 | 誤操作 | 不正アクセス | バグ等 | 設定ミス | 内部犯罪 | ワーム等 | 不正な情報持ち出し | 不明 | 目的外使用 | その他 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 領収証 | 0.754 | 1.000 | 0.132 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 紛失 | 4.323 | 3.529 | 1.316 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 保管 | 0.246 | 1.529 | 0.316 | 0.000 | 0.286 | 0.000 | 0.400 | 0.000 | 0.091 | 0.000 | 0.000 | 2.000 | 0.000 |
| 作業 | 2.185 | 0.353 | 2.763 | 0.000 | 0.286 | 0.000 | 0.200 | 0.000 | 0.000 | 0.000 | 1.000 | 0.000 | 1.000 |
| 盗難 | 0.123 | 0.000 | 2.921 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.333 | 0.000 | 0.000 | 0.000 |
| ガス | 1.800 | 1.118 | 2.026 | 0.000 | 0.429 | 0.000 | 0.600 | 0.000 | 0.000 | 0.667 | 0.200 | 1.500 | 0.000 |
| 用 | 0.862 | 0.765 | 2.447 | 0.133 | 0.000 | 0.000 | 0.600 | 0.000 | 0.091 | 0.000 | 0.000 | 0.000 | 0.000 |
| メール | 0.000 | 0.000 | 0.053 | 4.400 | 1.071 | 1.000 | 0.200 | 0.000 | 0.000 | 0.000 | 2.000 | 0.000 | 0.000 |
| 送信 | 0.000 | 0.000 | 0.053 | 3.200 | 0.500 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.200 | 0.000 | 0.000 |
| メールアドレス | 0.000 | 0.000 | 0.000 | 3.000 | 1.571 | 1.000 | 0.200 | 0.000 | 0.000 | 0.000 | 0.200 | 0.000 | 0.000 |
| 配信 | 0.000 | 0.000 | 0.000 | 1.200 | 0.071 | 0.000 | 0.200 | 1.000 | 0.000 | 0.000 | 0.200 | 0.000 | 0.000 |
| 利用 | 0.662 | 0.412 | 0.658 | 0.467 | 3.214 | 2.000 | 1.200 | 0.000 | 0.000 | 0.000 | 1.600 | 1.000 | 0.000 |
| ウェブサイト | 0.015 | 0.000 | 0.053 | 0.267 | 1.786 | 0.000 | 0.200 | 0.000 | 0.000 | 0.000 | 0.600 | 0.000 | 0.000 |
| 不正アクセス | 0.000 | 0.000 | 0.000 | 0.000 | 2.357 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 登録 | 0.092 | 0.000 | 0.079 | 0.467 | 1.143 | 12.000 | 1.200 | 0.000 | 0.000 | 0.000 | 0.000 | 2.500 | 2.000 |
| 会員 | 0.000 | 0.000 | 0.000 | 0.267 | 0.500 | 8.000 | 0.400 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 他 | 0.077 | 0.059 | 0.079 | 0.267 | 0.214 | 8.000 | 0.400 | 0.000 | 0.000 | 0.000 | 0.000 | 0.500 | 0.000 |
| 事象 | 0.277 | 0.235 | 0.211 | 0.600 | 0.357 | 8.000 | 0.800 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 2.000 |
| パスワード | 0.123 | 0.000 | 0.132 | 0.067 | 3.500 | 7.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 管理 | 0.785 | 1.235 | 0.816 | 0.933 | 0.286 | 3.000 | 3.600 | 0.000 | 1.273 | 2.667 | 1.800 | 0.000 | 0.000 |
| 画面 | 0.015 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 | 2.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 閲覧 | 0.108 | 0.059 | 0.079 | 0.467 | 0.500 | 5.000 | 2.600 | 0.000 | 0.000 | 0.000 | 0.400 | 0.000 | 0.000 |
| 外部 | 0.508 | 0.471 | 0.500 | 0.067 | 0.929 | 0.000 | 2.000 | 0.000 | 0.091 | 0.333 | 1.000 | 0.000 | 0.000 |
| ホールディングス | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.200 | 2.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 関係 | 0.431 | 0.647 | 0.474 | 1.067 | 0.357 | 0.000 | 1.200 | 2.000 | 0.909 | 1.667 | 0.400 | 1.500 | 0.000 |
| 女性 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.400 | 2.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 私 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.400 | 2.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| たち | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.400 | 2.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 徹底 | 0.462 | 0.294 | 0.500 | 1.067 | 0.143 | 1.000 | 0.600 | 0.000 | 1.364 | 2.000 | 0.600 | 0.000 | 2.000 |
| 上 | 0.138 | 0.235 | 0.658 | 0.067 | 0.643 | 1.000 | 1.200 | 0.000 | 1.273 | 1.000 | 0.400 | 1.000 | 0.000 |
| パソコン | 0.323 | 0.000 | 0.763 | 0.067 | 0.286 | 0.000 | 0.000 | 0.000 | 1.455 | 4.000 | 0.800 | 0.000 | 0.000 |
| 業務 | 1.108 | 0.294 | 2.526 | 0.067 | 0.071 | 0.000 | 0.600 | 0.000 | 0.909 | 3.333 | 1.800 | 0.000 | 0.000 |
| 委託 | 1.015 | 0.529 | 1.316 | 0.267 | 0.643 | 0.000 | 0.200 | 0.000 | 0.636 | 2.333 | 0.200 | 0.000 | 0.000 |
| 流出 | 0.462 | 0.471 | 0.632 | 0.533 | 2.214 | 0.000 | 1.200 | 0.000 | 2.364 | 3.333 | 3.800 | 0.000 | 0.000 |
| 攻撃 | 0.000 | 0.000 | 0.000 | 0.067 | 0.500 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 2.800 | 0.000 | 0.000 |
| 調査 | 0.400 | 0.353 | 0.079 | 0.000 | 1.929 | 2.000 | 0.400 | 0.000 | 0.273 | 0.667 | 3.400 | 0.500 | 1.000 |
| データ | 0.154 | 0.176 | 0.184 | 0.000 | 0.143 | 0.000 | 0.800 | 1.000 | 0.273 | 0.000 | 2.600 | 0.000 | 0.000 |
| 個人情報 | 0.677 | 0.588 | 1.184 | 0.933 | 0.571 | 4.000 | 3.200 | 0.000 | 1.000 | 1.333 | 2.200 | 0.000 | 0.000 |
| 者 | 0.277 | 0.706 | 0.421 | 1.200 | 0.643 | 1.000 | 0.800 | 0.000 | 0.364 | 0.667 | 1.000 | 12.000 | 2.000 |
| 事業 | 0.646 | 0.529 | 0.289 | 0.400 | 0.071 | 0.000 | 0.400 | 0.000 | 0.182 | 1.000 | 0.200 | 7.000 | 2.000 |
| 小売 | 0.000 | 0.000 | 0.000 | 0.267 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 5.500 | 0.000 |
| 建物 | 0.108 | 0.000 | 0.184 | 0.000 | 0.000 | 0.000 | 0.200 | 0.000 | 0.000 | 0.000 | 0.000 | 5.500 | 0.000 |
| 書類 | 1.431 | 1.588 | 1.026 | 0.067 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 6.000 | 0.000 |
| 社 | 0.138 | 0.059 | 0.079 | 0.067 | 0.000 | 0.000 | 0.200 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 11.000 |
| 電気 | 0.431 | 0.000 | 0.316 | 0.333 | 0.429 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 5.000 |
| 当該 | 0.785 | 1.471 | 0.395 | 0.467 | 1.857 | 2.000 | 0.400 | 1.000 | 0.455 | 0.000 | 1.600 | 2.000 | 5.000 |
| 確認 | 1.108 | 1.765 | 0.868 | 0.733 | 2.214 | 4.000 | 2.000 | 0.000 | 0.727 | 1.000 | 1.400 | 2.500 | 4.000 |
| 当社 | 0.938 | 1.588 | 1.263 | 1.600 | 0.500 | 1.000 | 1.000 | 0.000 | 1.727 | 4.333 | 1.800 | 0.000 | 3.000 |

表 3.2　調査対象企業の統計情報 (企業数)

| 業種 | 企業数 | ISMS | CIO | 大企業 2 | 大企業 1 | 中小企業 |
|---|---|---|---|---|---|---|
| 情報通信・サービスその他 | 104 | 14 | 17 | 17 | 43 | 44 |
| 素材・化学 | 80 | 27 | 23 | 31 | 37 | 12 |
| 食品 | 40 | 19 | 12 | 13 | 19 | 8 |
| 電機・精密 | 68 | 28 | 23 | 31 | 29 | 8 |
| 機械 | 56 | 15 | 15 | 15 | 33 | 8 |
| 建設・資材 | 79 | 25 | 27 | 36 | 34 | 9 |
| 金融 (除く銀行) | 1 | 0 | 0 | 0 | 0 | 1 |
| 小売 | 18 | 2 | 1 | 3 | 9 | 6 |
| 医薬品 | 19 | 10 | 7 | 14 | 2 | 3 |
| エネルギー資源 | 5 | 4 | 2 | 2 | 3 | 0 |
| 商社・卸売 | 18 | 2 | 6 | 1 | 7 | 10 |
| 不動産 | 4 | 0 | 2 | 0 | 0 | 4 |
| 自動車・輸送機 | 12 | 4 | 5 | 6 | 4 | 2 |
| 鉄鋼・非鉄 | 20 | 11 | 5 | 13 | 5 | 2 |
| 電機・ガス | 12 | 7 | 4 | 11 | 0 | 1 |
| 運輸・物流 | 1 | 1 | 1 | 0 | 0 | 1 |
| 合計 | 537 | 169 | 150 | 193 | 225 | 119 |

表 3.3　クローラーの実験結果

| 期間 | 企業数 | 取得記事数 | インシデント記事数 | 割合 |
|---|---|---|---|---|
| 2004/10/1 - 2018/11/2 | 537 | 17,957 | 191 | (0.01) |

表 3.4　推定結果

| | 日付 | 被害人数 | 漏洩原因 | 日付&規模&原因 |
|---|---|---|---|---|
| 適合率 | 0.882<br>157/178 | 0.792<br>141/178 | 0.719<br>128/178 | 0.505<br>90/178 |

# 第 4 章

# 取得したインシデント情報の評価

## 4.1　JNSA との比較

表 4.1 にインシデントを確認できた企業数とインシデント数を示す．比較対象として，JNSA データセットを使用する．公平に比較するために，表 4.1 では本調査結果を 2005 年から 2016 年の 34 企業，141 インシデントに限定する．企業数，インシデント数はどちらも JNSA の約 0.55 倍となった．また，クローラー独自で収集できたインシデントも，JNSA 独自の件数と比べると半分以下の結果となった．

JNSA 独自で収集できていて，本クローラーにより収集できなかったインシデント 171 件について，企業のウェブサイトを調査したところ，約 8 割にあたる 134 インシデントがウェブサイトに掲載されていないことを確認した．また 134 件のうち，インシデントの日付まで遡れないものが 40 件，日付は遡れるがウェブサイトに掲載がないものが 94 件であった．

図 reffig2.5 に JNSA との共通インシデントと，本調査独自のインシデント例を示す．企業のプレスリリースのみから取得できるインシデントは，人的ミスによるものなど事件性の低いものだけだと考えていたが，不正アクセスなど事件性の高いインシデントも含まれていた．

インシデント数の年による変化を図 4.2 に示す．JNSA のインシデントは本調査の調査対象であった 537 企業に限定している．2005 年から 2012 年にかけてのインシデント数は，JNSA よりも少ないが，これはインシデントについてのリリースが HP に存在しないことが原因である．

図 4.3 に，インシデントの例と JNSA データセットと自動分類システムにより抽出した項目を示す．インシデントリリースを入力として，抽出できた取得項目を出力する．正解の出力は JNSA のデータセットにある項目とした．検出結果には，正しい検出と誤検出があるが，この例では全て正しく検出できている．インシデント内容要約以下の項目は未実装である．

表 4.1　調査結果と JNSA との比較

| | JNSA | JNSA 独自 | 本調査 | 本調査独自 | 共通 |
|---|---|---|---|---|---|
| 企業数 | 65 | 42 | 34 | 28 | 23 |
| インシデント数 | 251 | 171 | 141 | 61 | 80 |

株式会社ディー・エヌ・エー

2016/04/01

DeNAが運営している「Mobage」において、第三者が「Mobage」ユーザに成りすまし、不正にログインしたと思われる事象が判明した。
不正ログインの確認されたID件数
最大104,847件

Jnsaとの共通インシデント

グリー株式会社

2016/12/27

グリーが運営するソーシャル・ネットワーキング・サービス（SNS）「GREE」およびスマートフォン向けゲームアプリを管理しているシステムに対して、不正なアクセスがあったことが判明

Crawler独自のインシデント

図 4.1　共通インシデントと独自インシデント例

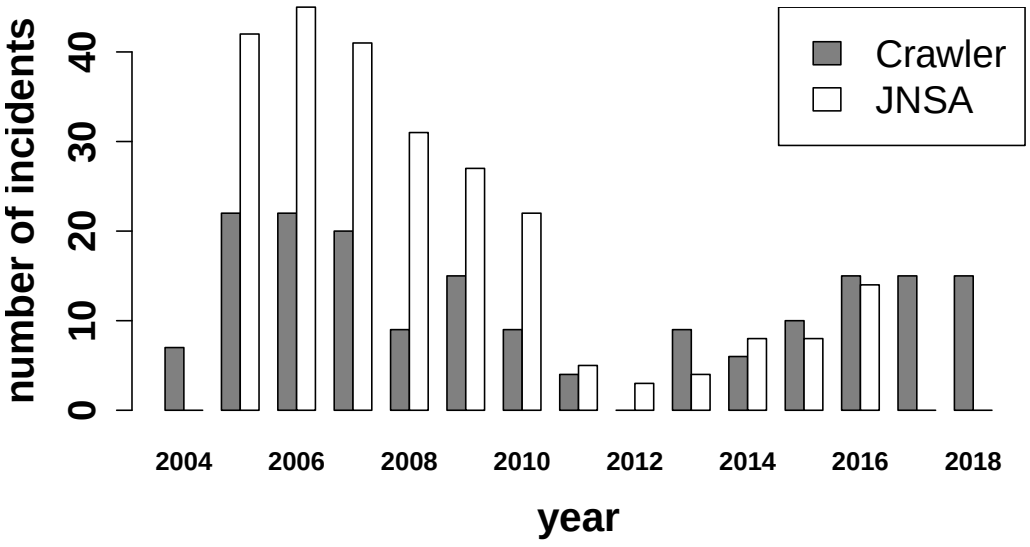


図 4.2　インシデント数経年変化

## 4.2　聞蔵 II での調査結果との比較

表 4.2 に [6] と JNSA，本調査の比較を示す．[6] の調査では 2015 年分のみ収集したので，比較対象が一年分となっている．また，[6] による調査では 279 件のインシデントを収集したが，本調査対象の 537 社に絞ると 3 件のみになった．すべてに共通したインシデントは 2 件のみで，クローラーによるインシデント数が最も多い．

インシデントリリース(入力)

株式会社ディー・エヌ・エー
2016/04/01
DeNAが運営している「Mobage」において、第三者が「Mobage」ユーザに成りすまし、不正にログインしたと思われる事象が判明した。不正ログインの確認されたID件数最大104,847件

取得項目(出力)

| | 抽出結果 | JNSA(正解) |
|---|---|---|
| 企業名 | ディー・エヌ・エー | ディー・エヌ・エー |
| 業種 | 情報通信・サービスその他 | 情報通信業 |
| 日付 | 2016/04/01 | 2016/04/01 |
| 被害人数 | 104847 | 104847 |
| 漏洩原因 | 不正アクセス | 不正アクセス |
| インシデント内容要約 | 未定義 | 有 |
| 参照記事のURL | | |
| 社会的責任度 | | 普通 |
| 漏洩情報区分 | | 個人情報 |
| 漏洩経路 | | インターネット |
| 事後対応 | | 普通 |

図 4.3　プレスリリースからの取得項目の違い

表 4.2　三つのデータセットのインシデント数の比較

| | 本調査 | JNSA | 聞蔵 II | 共通 |
|---|---|---|---|---|
| 2015 | 10 | 8 | 3 | 2 |

## 4.3　考察

インシデントの自動分類は適合率 50% 以下であった．これは，各要素の適合率で最も低かった漏洩原因の誤推定が原因だと考えられる．表 4.3 に判定の結果を示す．紛失・置忘れの誤推定のうち 75% が管理ミスと判定された．これは紛失・置忘れと管理ミスの tf-idf により抽出した特徴語が複数被っていたことが原因だと考えられる．また，日付の誤判定は，日付を最新のものを出力しているが一部プログラムのバグにより最新の日付が抽出でてきていないことが原因だった．被害人数の誤判定 37 件のうち，被害人数自体を抽出できていなかったものが 12 件，残りの 25 件は被害人数を出力しているが正解の値とは異なっていた．25 件の内，実際の被害人数よりも大きな値を出力をしていたのが 15 件であった．被害人数誤抽出の例を図 4.4 に示す．図 4.4 のような例があるため，正規表現だけで正しい被害人数を抽出するのは限界があると考えられる．

JNSA データセットとの比較の結果，本調査の 537 企業におけるインシデント数は JNSA の約半分であった．しかし，収集できていないインシデントの約 5.5 割は HP 上に存在せず，約 2.5 割は期間が古いため確認できなかった．2013 年，2015 年，2016 年のように近年のインシデントについては JNSA よりも多く収集できていた．

表 4.3　各漏洩原因の識別数

| | 紛失・置忘れ | 管理ミス | 盗難 | 誤操作 | 不正アクセス | バグ等 | 設定ミス | 内部犯罪 | ワーム等 | 不正な情報持ち出し | 不明 | 目的外使用 | その他 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 紛失・置忘れ | 53 | 9 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 |
| 管理ミス | 5 | 11 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 盗難 | 3 | 2 | 29 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 1 | 0 | 0 |
| 誤操作 | 0 | 0 | 0 | 12 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 不正アクセス | 0 | 0 | 0 | 1 | 10 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| バグ・セキュリティホール | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 設定ミス | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 1 | 0 | 0 |
| 内部犯罪・内部不正行為 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| ワーム・ウイルス | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 5 | 3 | 1 | 0 | 0 |
| 不正な情報持ち出し | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 |
| 不明 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 |
| 目的外使用 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| その他 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

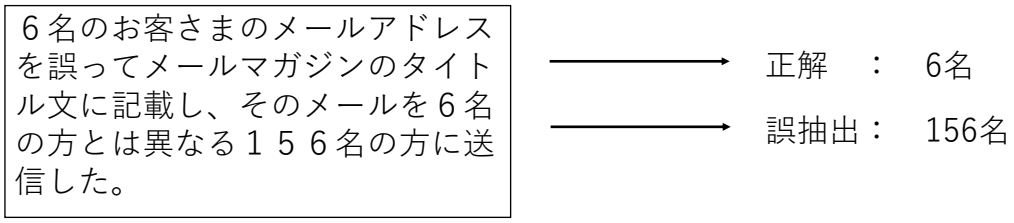


図 4.4　被害人数語抽出の例

# 第 5 章

# セキュリティマネジメントの効果

本稿では，提案データセットの効果を確認するため，本調査によるデータセットと CSR データセットの突合を行い，ロジスティック回帰を行なった．本調査と [8] による結果を表 5.1 に示す．正の係数は（業種に当てはまる時，該当年の時，マネジメントを実施している）インシデントの生起確率が増加することを表す．[8] では，業種や企業規模などの交絡因子を除くと 6 個のうち 5 個のマネジメントでインシデント生起確率が減少することを示した．本調査では 6 個中 3 個のマネジメントの係数が負になった．また，[8] では外部監査が正であったが本調査では負となった．これらの違いは [8] と比べて，調査企業数が少なくインシデント数が異なることが原因だと考えられる．

表 5.1　本調査によるデータセットを用いたロジスティック回帰

| | | Estimate | Std.E*RR*or | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| *a* | (Intercept) | −23.260 | 2084.000 | 0.991 | |
| *b* | 建設. 資材 | 16.280 | 2084.000 | 0.994 | |
| | 素材・化学 | 16.950 | 2084.000 | 0.994 | |
| | 自動車・輸送機 | −0.279 | 4188.000 | 1.000 | |
| | 鉄鋼・非鉄 | −0.012 | 3548.000 | 1.000 | |
| | 電機・精密 | 16.260 | 2084.000 | 0.994 | |
| | 情報通信・サービスその他 | 18.120 | 2084.000 | 0.993 | |
| | 電気・ガス | 20.330 | 2084.000 | 0.992 | |
| | 運輸・物流 | 0.367 | 14320.000 | 1.000 | |
| | 商社・卸売 | 0.467 | 3777.000 | 1.000 | |
| | 小売 | 17.540 | 2084.000 | 0.993 | |
| | 金融（除く銀行） | 1.145 | 14510.000 | 1.000 | |
| | 機械 | −0.219 | 0.065 | 1.000 | |
| *c* | 2014 | −0.350 | 0.724 | 0.629 | |
| | 2015 | −0.763 | 0.784 | 0.330 | |
| | 2016 | 0.752 | 0.595 | 0.206 | |
| | 2017 | 1.186 | 0.706 | 0.093 | . |
| *d* | LOG(従業員数) | 0.399 | 0.366 | 0.275 | |
| *x* | ISMS | 1.200 | 0.674 | 0.075 | . |
| | CIO | 0.000 | 0.719 | 1.000 | |
| | 内部監査 | 2.429 | 1.816 | 0.181 | |
| | 外部監査 | −0.959 | 0.428 | 0.025 | * |
| | 内部窓口 | −1.717 | 1.823 | 0.346 | |
| | 外部窓口 | −0.969 | 0.789 | 0.220 | |

表 5.2　[8] でのロジスティック回帰

| | | Estimate | Std.E*RR*or | Pr(>\|z\|) |
|---|---|---|---|---|
| *a* | (Intercept) | −8.300 | 1.072 | 0.000 *** |
| *b* | 建設・資材 | 0.223 | 0.800 | 0.780 |
| | 素材・化学 | −0.046 | 0.775 | 0.952 |
| | 自動車. 輸送機 | −0.334 | 0.981 | 0.734 |
| | 鋼鉄・非鉄 | −0.838 | 1.325 | 0.527 |
| | 電機・精密 | 0.091 | 0.805 | 0.910 |
| | 情報通信・サービスその他 | 0.561 | 0.738 | 0.448 |
| | 電気・ガス | 2.436 | 0.916 | 0.008 *** |
| | 運輸・物流 | 0.829 | 0.854 | 0.332 |
| | 商社・卸売 | 0.066 | 0.849 | 0.938 |
| | 小売 | 0.904 | 0.756 | 0.231 |
| | 金融（除く銀行） | 0.209 | 0.913 | 0.819 |
| | 機械 | −0.219 | 0.921 | 0.812 |
| *c* | 2014 | 0.221 | 0.333 | 0.507 |
| | 2015 | 0.185 | 0.343 | 0.590 |
| | 2016 | 0.185 | 0.350 | 0.597 |
| | 2017 | −0.193 | 0.374 | 0.607 |
| *d* | LOG(従業員数) | 0.948 | 0.255 | 0.000 *** |
| *x* | ISMS | −0.217 | 0.331 | 0.513 |
| | CIO | −1.097 | 0.330 | 0.001 *** |
| | 内部監査 | −0.207 | 0.374 | 0.580 |
| | 外部監査 | 0.117 | 0.277 | 0.674 |
| | 内部窓口 | −0.050 | 0.761 | 0.947 |
| | 外部窓口 | −0.685 | 0.296 | 0.021 ** |

# 第 6 章

# おわりに

本研究では，537 企業のプレスリリースについてクローリングを行い．171 件のインシデントを収集し，JNSA，「聞蔵 Ⅱ」によるデータセットとの比較を行った．クローリングでは取得記事のうちインシデントは 1 割であった．収集したインシデントと JNSA データセットの比較結果，61 件の独自のインシデントが見つかった．インシデントの日付，被害人数，漏洩原因は 70% 以上の適合率で正しい分類ができたが，全要素を正しく分類する例は 50% 以下であった．

今後は，インシデンの識別精度を高め，対象の企業数を増やし汎用性の高い DB を提供することを課題とする．

# 謝辞

本研究を遂行するにあたり，インシデントデータを提供いただいた日本ネットワークセキュリティ協会様に感謝いたします．

本研究では明治大学総合数理学部現象数理学科，乾孝治教授が受けている JSPS 科研費 JP16K03755 で購入した CSR データセットを使用しました．乾孝治教授に深く感謝いたします．

本研究を行うにあたり，多くの方より御指導いただきました．明治大学総合数理学部先端メディアサイエンス学科，菊池浩明教授に深く感謝申し上げます．予備実験等に協力してくださった菊池研究室の皆様並びに先端メディアサイエンス学科の方々に深く感謝の意を表するとともに，謝辞とさせていただきます．

# 参考文献


[1] NPO 日本ネットワークセキュリティ協会セキュリティ被害調査ワーキンググループ, 長崎県立大学情報システム学部情報セキュリティ学科, “2016 年 情報セキュリティインシデントに関する調査報告書個人情報漏洩編”, 2017.

[2] 山田道洋，菊池浩明，松山直樹，乾考治，個人情報漏洩の損害額の新しい数理モデルの提案， 第 82 回 CSEC 研究会, No.19, pp.1-6, 2018

[3] 情報セキュリティインシデント調査報告書（JNSA データセット）

[4] 聞蔵 II，朝日新聞社，オンライン記事データベース

[5] 豊田秀樹，「データマイニング入門-R で学ぶ最新　　データ分析-」，東京図書，pp.147-183，2008.

[6] 池上和輝，菊池浩明，インシデント調査に基づく漏えい原因のデータマイニング, 情報処理学会第 80 回大会，2W-06， vol.3，pp.543-544，2018.

[7] 東洋経済データサービス CSR データ

[8] 山田道洋，池上和輝，菊池浩明，乾考治，経営マネジメント状況による情報漏洩インシデント削減効果の評価 (2)，Computer Security Symposium 2018, pp.376-384, 2018.