

人工知能を活用した犯罪・交通事故  
発生予測技法の調査研究

補遺

(犯罪発生予測編)

平成 31 年 3 月

# 目次

1	実証実験の実施概要	
(1)	実証実験について	1
(2)	人工知能について	3
(3)	予測結果の評価方法について	5
(4)	特徴量と利用データ	9
2	実証実験結果	
(1)	第1回実験	14
(2)	第2回実験	18
(3)	第3回実験	28
3	X A Iについて	33
○	用語集	37
○	注釈	41
○	参考文献	42
○	別添1 SHAPによる人工知能の学習傾向分析結果	
○	別添2 特徴量一覧	

## 【実証実験の対象罪種】

本調査研究を行うに当たり、実証実験の対象罪種は次のとおりとする。

### ○ 空き巣

刑法第 235 条に規定する窃盗罪のうち、家人等が不在の住宅の屋内に侵入し、金品を窃取するもの（未遂を含む。）をいう。

### ○ 振り込め詐欺前兆電話

当調査研究においては、振り込め詐欺等で金品をだまし取るために前準備として、息子や孫、警察官等を装って架かってくる電話をいう。

### ○ 女性・子どもを対象とした犯罪・前兆事案

本調査研究においては、子供及び女性を対象とする性犯罪等（子供の生命又は身体を害する犯罪及び女性に対する性的犯罪（犯罪手口資料取扱規則（昭和 57 年国家公安委員会規則第 1 号）第 3 条第 8 号に規定する性的犯罪をいう。）をいう。）及びその前兆とみられる事案（声かけ、つきまとい、不審者のはいかい、公然わいせつ、痴漢、卑わい行為等の性犯罪に発展するおそれのある事案）をいう。

### ○ ひったくり

刑法第 235 条に規定する窃盗罪のうち、携帯している金品をひったくって窃取するもの（未遂を含む。）をいう。

### ○ 自転車盗

刑法第 235 条に規定する窃盗罪のうち、自転車を窃取するもの（未遂を含む。）をいう。

## 1 実証実験の実施概要

### (1) 実証実験について

人工知能を用いて犯罪発生予測を行う実証実験を3回に分けて実施した。

#### ア 第1回実験

単一の機械学習アルゴリズム、必要最低限と考えられる特徴量、神奈川県内全エリアを予測対象として、簡素な実験設定で予測実験を行った。

第2回目以降に実験する人工知能による犯罪発生予測のイメージの共有及び課題の洗い出しを行うことを目的とした。

結果は委員会において犯罪発生予測のイメージを共有した。また、課題としては、特徴量不足や予測対象範囲の絞込みなどが挙げられた。

#### イ 第2回実験

第1回実験の結果を踏まえ、オープンデータ等から入手可能な特徴量を追加するとともに、予測対象エリアを絞った。この条件の下、複数の機械学習アルゴリズム間の精度比較を行うことを目的とした。

複数の機械学習アルゴリズムの精度比較の結果、予測精度の高さに基づいて機械学習アルゴリズムの絞込みを行った。

#### ウ 第3回実験

第2回実験で予測精度が高かったアルゴリズムで実証実験を行った。特徴量の見直し、絞り込みを行った状態でも予測精度を維持できるかを確認することを目的とした。

結果はわずかな予測精度低下で抑えられ、予測対象に関係のある特徴量の重要度の傾向が分かりやすくなった。

◆ 図表1-1 実証実験ワークフロー



No.	プロセス名	実施概要
1	課題認識	実証実験の目的に認識齟齬がないことを委員会において確認する。
2	目標設定	委員とともに課題に対する業務活用目標を設定する。また、目標を達成するための精度目標も設定する。その他分析対象の罪種の選定等も行う。
3	データ提供	実証実験に必要な業務データ等を神奈川県警察から委託事業者へ提供する。
4	データ取得	提供されたデータやオープンデータ等を実証実験環境へ移行する。
5	データ把握	データ特性の把握を行う。主に基本統計分析を行い、データが人工知能の特徴量として利用可能かどうかを検証する。
6	仮説立案(特徴量設計)	データ把握で取得した情報を基に業務活用目標に対し、どのようなデータを人工知能に学習させると目標を達成できるかという仮説を立案する。立案した仮説を基に特徴量の設計を行う。
7	データ準備	仮説及びデータの特徴を考慮し、学習用データ(特徴量)を作成する。
8	学習(モデル化)	学習用データを用いて機械学習を行い、予測モデルを作成する。
9	予測	予測モデルの精度を評価するため、検証用データを用いて予測を行う。
10	予測結果分析	機械学習でよく使われる評価指標に基づいて、モデル別に予測結果の分析、評価を実施。分析結果に基づいて、精度向上案を委員会において検討する。

## (2) 人工知能について

本調査研究における「人工知能」とは、「機械学習を行うアルゴリズム」と定義した。

複数の機械学習アルゴリズムを用いて犯罪発生予測の実験を行い、その結果を比較することで、犯罪予測に適した機械学習アルゴリズムを探ることが実験の目的の一つである。

本実証実験において使用する機械学習アルゴリズムは、インターネット上にライブラリとしてプログラムのソースコードが公開され、誰もが無償で利用可能なオープンソースを使用した。

オープンソースを使用する理由は

- アルゴリズムの仕組みやプログラムが広く公開されていることから、アルゴリズムを比較評価する公平性や透明性を確保できること。
- 世界中のユーザーによりバグ修正や改良が進められているため、動作安定性が高く、利用可能なライブラリ等も豊富であること。
- Kaggle<sup>(1)</sup>で開催される予測コンペなどで、オープンソースが用いられ、高い予測性能を上げており、それらの結果を活かした精度向上の可能性を秘めていること。

などの観点からである。

実証実験で使用するオープンソースは、以下の4つである。

括弧内は、アルゴリズムの分類を示す。

- ・ ロジスティック回帰（確率的線形判別）：scikit-learn<sup>(2)</sup>
- ・ XGBoost（決定木+アンサンブル学習）：PyXGBoost<sup>(3)</sup>
- ・ サポートベクターマシン（カーネル法）：scikit-learn

・ディープラーニング（多層ニューラルネットワーク）：

TensorFlow<sup>(4)</sup>、Keras<sup>(5)</sup>

この4種類のアプローチを選定した理由については図表1-2のとおりとする。

◆ 図表1-2 アルゴリズムの特徴と採用理由

	アルゴリズム	特徴	選択理由
1	ロジスティック回帰	<ul style="list-style-type: none"> <li>代表的な機械学習アルゴリズムの一つ</li> <li>商品を購入するか否か、といったクラス分類の予測をする時に用いられる</li> <li>設定するハイパーパラメータが少なく、計算も速い</li> <li>複数のアルゴリズムによる精度比較の際のベースとして用いられることが多い</li> </ul>	<p>様々な予測問題において扱われてきた実績がある。また、シンプルな線形回帰であるため、学習したモデルの解釈を行いやすいというメリットがある。比較のためのベースとして採用する。</p>
2	XGBoost	<ul style="list-style-type: none"> <li>決定木と呼ばれる木構造のルールをいくつも作成して予測するアルゴリズム</li> <li>予測値が外れたときに、その誤差をさらに埋めるよう学習が進む</li> <li>アンサンブル学習と呼ばれる方法で高い予測精度を誇るが、設定すべきハイパーパラメータが多い</li> </ul>	<p>世界的なデータ解析コンペティションで高い精度を出したことで有名になり、さまざまな予測に活用されている。学習モデルが予測に際して重視している特徴量を確認することが可能である。犯罪予測にも適しているアルゴリズムと考えられるため採用する。</p>
3	サポートベクターマシン	<ul style="list-style-type: none"> <li>代表的な機械学習アルゴリズムの一つ</li> <li>カーネルと呼ばれる方法で非線形分離が可能</li> <li>少ないデータでも高い予測精度を期待できるが、データが大量であると計算時間が膨大化する</li> </ul>	<p>ロジスティック回帰よりも複雑な問題を解くことに長けている。学習モデルの解釈は難しいが予想精度は一般的に高い傾向があるため採用する。</p>
4	ディープラーニング	<ul style="list-style-type: none"> <li>人間の脳神経回路をモデルにした多層構造アルゴリズム</li> <li>画像、音声、自然言語の分野で高い認識精度を誇る</li> <li>一般的に大量のデータが必要である場合が多く、また学習した結果を人間が解釈することが難しいため、世界中で更なる研究が進められている</li> </ul>	<p>近年の人工知能ブームの中心にあり、画像や音声認識で高い予測精度を発揮していることから、犯罪発生予測に適用できるか検討するために採用する。</p>

### (3) 予測結果の評価方法について

人工知能の予測性能の評価については、様々な評価指標が使われている。

犯罪予測を例として説明すると、予測結果と実際の結果は以下のような混合行列となり、これをクロス集計して評価することになる。

◆ 図表1-3 混合行列

	「発生する」と予測	「発生しない」と予測
実際に発生	真陽性(TP)	偽陰性(FN)
実際に未発生	偽陽性(FP)	真陰性(TN)

この4つの数値 (TP,FP,FN,TN) を用いて、一般的に使われている評価指標を挙げる。

- ・ 正解率 (Accuracy)

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

「発生する」「発生しない」を正しく予測した割合。

- ・ 適合率 (Precision)

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

「発生する」と予測して実際に発生した割合。

- ・ 再現率 (Recall)

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

実際に発生した場所を「発生する」と予測できた割合。



・ F 値 (F-measure)

$$F\text{-measure} = (2 \times \text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

適合率と再現率の調和平均。トレードオフとなる双方の評価値のバランスの良さを表す総合的評価指標といえる。

・ リスク比

$$\text{Risk Ratio} = (\text{TP} / (\text{TP} + \text{FP})) / (\text{FN} / (\text{TN} + \text{FN}))$$

人工知能の能力評価指標ではないが、発生予測場所と未発生予測場所における発生率の比。

「発生する」と予測した場所が「発生しない」と予測した場所よりも犯罪が起きるリスクが何倍高いかを表す指標であり、全くランダムに予測するモデルである場合のリスク比は1程度になる。

理論上は無限大までの数値を取ることがある。

どの評価指標を使用して予測モデルの評価をするかということが問題であるが、適用する対象やどのような予測結果を使用するかで重視する評価指標を使い分ける必要がある。

一見すると、「正解率」が高い予測モデルが高性能であるという印象を受けるかも知れないが、データに偏りがある場合には注意する必要がある。

日本国内における犯罪の現状からいえば、発生件数は諸外国と比較して少なく、全体で見ると犯罪が発生することはまれな事象である。

予測単位の設定にも大きく依存するが、発生データ・未発生データは通常は未発生側に大きく偏っていると考えられる。

正解率に関していえば、算出式を見れば分かるとおり、データの偏りが大きければ大きいほど、偏っている側の予測（真陰性予測）をすれば正解率が相当に高くなってしまうため、犯罪予測に関していえば全て「発生しない」という予測をするだけで正解率が99%を超える数値になってしまう。

実際には何の予測もできないモデルが非常に高い正解率を出してしまうこともあるため、犯罪予測の評価指標としては適さない（100%にならない限り犯罪予測の評価指標としては大きな意味はない）と考えられる。

犯罪予測に関していえば、「発生する」という予測が

- ・ 正確な予測か
- ・ 漏れがない予測か

という部分が最も重要な要素であるといえる。

発生予測の正確性は適合率で、網羅性は再現率で判断することができるが、適合率と再現率は片方の指標の値が上昇すると片方の指標の値が下降するトレードオフの関係であるため、どちらを重視するか、どの程度のバランスを取るかを考える必要がある。

適合率を上げることで、真陽性予測の正確性が重視されるようになるが、計算式を見てもわかるように偽陰性予測を考慮しない指標であるため、偽陰性予測が増えてしまう可能性が高まる。

一方、再現率を上げると犯罪発生の取りこぼしを減らすことが可能であるが、偽陽性予測を考慮しない指標であるため、偽陽性予測が増えてしまうことも考えられる。

この2つの評価指標の双方のバランスを重視した指標がF値である。

F値を用いることで、適合率と再現率というトレードオフになる2つの評価指標を、統合して一つの指標として評価できることが利点である。

アルゴリズムやモデルの比較評価を行う場合は、F値を利用することで総合的な性能評価が可能であると考えられる。

人工知能を用いた犯罪予測に関していえば、適合率と再現率、その総合的な指標であるF値を確認し、どの程度の評価値であれば、実業務に活用が可能なのを見極めながら検討する必要があると考えられる。

#### (4) 特徴量と利用データ

犯罪発生データとして

- ・ 刑法犯認知情報
- ・ 犯罪事件管理システム情報
- ・ 振り込め詐欺前兆電話入電状況データ

を使用した。

刑法犯認知情報に関しては警察庁の統一の内容のものから、発生日時、発生場所、罪名、手口などの基本的な項目を抽出して使用した。

刑法犯認知情報を使用した予測対象罪種は

- ・ 空き巣
- ・ ひったくり
- ・ 自転車盗

である。

犯罪事件管理システム情報は、神奈川県警察で運用している犯罪事件管理システムに蓄積された事件・事案に関する情報である。

基本的には刑法犯認知情報と同様の内容が記録されているが、それに加えて事件化しなかった取扱い（事案処理）についても措置情報が入力されていることから、「女性・子どもを対象とした犯罪・前兆事案」については犯罪事件管理システム情報からデータを抽出して使用した。

振り込め詐欺前兆電話入電状況データは、振り込め詐欺などの特殊詐欺のだましの電話、いわゆる「前兆電話」「アポ電話」などと呼ばれているものであり、住民から警察署に情報提供されたものを集計し、CSVファイルとして入力したデータである。

振り込め詐欺前兆電話入電状況データからは、入電時刻・場所に関する情報を使用した。

これらのデータを人工知能で処理が可能となるように、以下のデータ加工処理を行った。データ加工処理は神奈川県警察が行う作業と委託業者が行う作業の役割分担が必要であった。

例えば、「住所や氏名」のような個人情報のデータが入っているデータの加工処理は、個人情報保護の観点から神奈川県警察側でしかできない作業であり、委託業者側の作業の例としては位置座標からメッシュコードへの変換、日時のデータクレンジングやデータの特徴量化などである。

#### ア 住所や氏名

氏名などの個人情報に関する情報のマスクや住所と施設名・店舗名・住宅名を分離し、データの入力間違い、誤記等を更に修正した。

修正した住所に対して、アドレスマッチング（ジオコーディング）を行い、位置座標を付与した。

アドレスマッチングには、東京大学空間情報科学研究センターの「CSVアドレスマッチングサービス」<sup>⑥</sup>を利用し、全国街区表示レベルでアドレスマッチングを行った。

マッチングレベルが「丁目」より粗いもの（位置座標の合致精度が低いもの）については、神奈川県警察の統合地図情報管理システムから手作業で地点を割り出し、位置座標情報を取得して付与した。

ジオコーディングにより得られた位置座標を、100メートル分割メッシュのメッシュコードに変換を行い、メッシュの犯罪発生情報とした。

## イ 日時

日時についてはデータが一部欠損しているものについて以下のデータ加工処理を行った。

- 年・月が不明なデータは、原則として学習データから除外した。

(例：平成 30 年以下不明)

- 発生の自至に時間幅があり、日にちが不明なものについては一律に発生日付を「1日」として扱うことにした。

(例：平成 30 年 7 月以下不明→平成 30 年 7 月 1 日)

- 時間幅がなく、日付不明のものについては学習データから除外した。

- 発生時刻について不明なものは、時間幅があるものについては「0時」、「0分」として、時間幅がないものについては除外した。

(例：1月1日午後8時以下不明→1月1日午後8時00分)

空き巣や自転車盗などの非面接の窃盗犯罪に言えることであるが、発生日時が「被害に遭っていないことが明らかである時点（始期）から被害に遭ったことに気が付いた時点（終期）まで」の時間幅を持っており、実際に犯罪が敢行された時刻が特定できない、という問題である。

例えば当県の空き巣の認知データで時間幅を見てみると、最短では数十分程度から長いものであると数年単位の幅があるものまで確認できた。

これらを何らかの方法で一意に定めなければ、その犯罪発生がどの時間帯に属するのかが決めることができない。

この問題に対して考えられる解決方法としては

- 予測単位時間（本実証実験でいえば3時間）内に収まる発生データのみを用いる。

- 検挙後の被疑者への調査データなどから、実際の発生時刻の分布を割り出し、その分布に従って発生時刻を割り当てる。
- 始期・終期のどちらか、又はその中間点を発生時刻とする。
- 始期と終期の間のランダムに割り振った1点を発生時刻とする。
- 始期と終期の間の全ての時間帯に発生及びその重みを均等に割り振る。

などの方法が考えられる。

予測単位時間内に収まるデータのみ限定すると、学習データが著しく減少してしまう（例えば空き巣に関して3時間以内の時間幅のものは、全体の1割強でしかない。）という問題が生じる。また、検挙後に被疑者の面接調査を行い、その発生分布をまとめたデータは存在しないことから、この方法についても現実的には困難である。

海外における先行研究では、ある1点に特定せず、始期・終期間の発生可能性のある時間帯に均等に発生重みを割り振ることが最も性能が高かったとする報告があるが、学習モデルが複雑化し、二値分類問題として予測することができなくなってしまうという問題がある。

前述の研究によると、発生始期、終期の間の1点をランダムに発生時刻として割り当てる方法は、始期、終期も又は中間点に割り当てるよりは高い分析性能を示したと報告されている。

本調査研究においては、発生時間に幅があるものについては、方法が単純で分かりやすい、「ランダム割当て」を用いて発生時刻を定める方法を採用することとするが、余りにも時間幅が大きいもの（72時間以上）はノイズとなるおそれがあることから、学習データから除外することとした。

この発生時刻決定方法の評価のため、空き巣の予測に関しては、他の罪種と同じ「3時間単位の予測モデル」と「24時間単位の予測モデル」の両方で予測実験を行い、予測精度を比較することでこの手法の適否を判断することとした。



## 2 実証実験結果

### (1) 第1回実験

空き巣及び振り込め詐欺前兆電話(オレオレ詐欺のみ)の発生予測を行った。

実験設定は、以下のとおりである。

#### ア 実験の設定

##### (ア) 機械学習アルゴリズム

使用する機械学習アルゴリズムは、ロジスティック回帰の1種類のみを用いた。

ロジスティック回帰を用いる理由としては

- 処理速度が高速である。
- 学習したモデルの解釈が行いやすい。

などである。

過学習(オーバーフィッティング)対策としては、交差検証(クロスバリデーション)やL1正則化により対応した。

##### (イ) 特徴量

犯罪等の発生に関すると考えられる、日時・場所、天候、人口、主な最寄り施設までの距離等を特徴量として使用した。

使用した特徴量は、別添2「特徴量一覧」の第1回実証実験を参照。

## (ウ) 対象期間

学習及び予測の対象とした期間は、以下のとおりである。

◆ 図表 2-1 第1回実験学習設定

業務種別	学習期間	正例学習件数	予測評価期間
空き巣	2010年1月～2018年2月	18,436	2018年3月
振り込め詐欺前兆電話	2013年1月～2018年2月	36,434	2018年3月

※ 第1回実験で設定した振り込め詐欺前兆電話はオレオレ詐欺のみ

それぞれのデータの最新1ヶ月分を除いた期間を学習データとして機械学習アルゴリズムに学習させることで得られた予測モデルにより最新の1ヶ月分を予測させて、実際の最新1ヶ月分と比較することにより予測の評価を行った。

学習データは、学習期間内の犯罪が発生したメッシュの状況を正例、犯罪が発生しなかったメッシュの状況を負例とした。

犯罪に関して言えば、正例及び負例の比率に大きく偏りがあり（負例が圧倒的に多い。）、そのまま学習させると予測に偏りが生じる（負例が多いまま学習すると、負例を予測することで正解率が上がることから、そのような予測結果を出す方向に学習が進む。）ため、負例については負例の全体から正例よりやや多い（約1.2倍）件数をランダムに選択して減じた上で、学習時に正例と同数に調整する処理（アンダーサンプリング）を行い、学習させることとした。

## (エ) 予測対象メッシュ

予測評価メッシュ数：32,041,600

内訳：神奈川県内総メッシュ数 258,400 × 予測対象日数 31 日 × 1 日当たりの  
時間分割数 4 = 32,041,600

予測の時間単位は第 1 回実験に限り、特徴量の設計や学習時間の関係から、6  
時間単位で行った。

予測の判断を分ける<sup>しきい</sup> 閾値については、得られた発生傾向確率が 50%以上で  
あれば「発生する」という予測結果、50%未満であれば「発生しない」という予  
測結果として扱い、予測結果と評価データを比較し、「予測が当たったかどう  
か」を示す評価指標を算出した。

### イ 実験結果

第 1 回予測実験の結果は、以下のとおりである。

◆ 図表 2-2 第 1 回実験予測結果

罪種	TP	FP	FN	TN	正解率	適合率	再現率	リスク比	F値
空き巣	103	13,142,963	4	18,898,530	58.98%	0.0008%	96.26%	37.03	0.0016%
振り込め詐欺前兆電話	477	12,530,112	1	19,511,010	60.89%	0.0038%	99.79%	742.72	0.0076%

再現率が極めて高く、実際に空き巣や振り込め詐欺架電が発生した場所は、ほ  
ぼ網羅して予測できていた。

その一方で、再現率とトレードオフの関係となる適合率が極めて低い結果とな  
った。

空き巣で言えば、発生すると予測した場所を全てカバーすればほぼ全ての発生  
を捕捉できるが、発生を予測したメッシュで実際に犯罪が発生する確率は  
0.0008%と極めて低く、偽陽性メッシュが大多数を占めるという予測結果となっ  
た。

全体的に偽陽性予測（FP）に偏っているため、発生の取りこぼしは少ないものの、実務に活用しようとする、対応に必要なリソース（警察官）が足りず、また、予測が外れて実際には発生しない場所が多いため（神奈川県全体の約4割）、警察力の効率的な配置を行うという目的には沿わない予測結果となった。

このような結果となった理由としては使用した特徴量が少なかったことが考えられる。

ロジスティック回帰は、学習で得られた予測モデル（回帰式）から予測に寄与する特徴量が分かるため、これを確認したところ

- ・時間帯
- ・土地利用区分

などのメッシュ間で差が出にくいものが重視されていた。

このことから、メッシュ間で差が出やすい特徴量が不足していることにより犯罪発生メッシュと未発生メッシュの特徴の差異を人工知能が捉えきれず、予測結果が偏る結果となったことが考えられた。

## (2) 第2回実験

第1回実験の結果を踏まえ、

- ・複数の機械学習アルゴリズムの適用
- ・予測対象罪種の追加
- ・特徴量の追加
- ・対象メッシュの絞り込み

などを行い、予測精度の変化を確認することを目的として第2回の実験を行った。  
実験設定は、以下のとおりである。

### ア 実験の設定

#### (ア) 機械学習アルゴリズム

同一条件で前述の4種の機械学習アルゴリズムを用いて学習と予測を行い、その予測性能を比較する。

#### (イ) 特徴量

第1回実験と比較して大幅に特徴量を追加した。追加した特徴量は警察業務データ、オープンデータなどのデータから作成した。

使用した特徴量は、別添2「特徴量一覧」の第2回実証実験を参照。

#### (ウ) 予測対象罪種

予測対象罪種については、第1回実験の「空き巣」「振り込め詐欺前兆電話(オレオレ詐欺のみ)」に加え、「女性・子ども(15歳未満)を対象とした犯罪・前兆事案」「女性・子ども(15歳以上)を対象とした犯罪・前兆事案」「ひったくり」「自転車盗」を対象罪種に加えた。また、予測対象罪種を検討する過程において「振り込め詐欺前兆電話」については、オレオレ詐欺のみだけでなく、還付金等詐欺、架空請求詐欺、融資保証金詐欺といった振り込め詐欺前兆電話に分類される手口を全て含めることとした。

## (エ) 対象期間

学習及び予測の対象とした期間は、以下のとおりである。

◆ 図表2-3 第2回実験学習設定

業務種別	学習期間	正例学習 件数	負例学習 件数	予測評価期間
空き巣	2011年1月～2017年6月	10,307	12,368	2017年7月～2018年6月
振り込め詐欺前兆電話	2014年1月～2016年12月	30,263	40,376	2017年1月～2017年12月
女性・子ども（15歳未満）を 対象とした犯罪・前兆事案	2014年1月～2016年12月	2,379	2,763	2017年1月～2017年12月
女性・子ども（15歳以上）を 対象とした犯罪・前兆事案	2014年1月～2016年12月	6,380	7,599	2017年1月～2017年12月
ひったくり	2011年1月～2017年6月	4,667	5,959	2017年7月～2018年6月
自転車盗	2011年1月～2017年6月	89,892	107,474	2017年7月～2018年6月

それぞれのデータの初期1年分及び最新1年分を除いた期間を学習データとして使用し（初期1年分には過去発生に関係する特徴量などとして利用）、最新1年間の評価期間の中から各月ごとランダムに3日又は4日を選択した計40日を予測評価期間として使用した。

1年間からランダムに40日を抽出している理由は、警察実務の経験測的に季節による犯罪発生件数の増減を体感することがあることや交通事故では発生の季節性が指摘されている<sup>2)</sup>ことから犯罪予測についても季節による影響がある可能性を考慮したためである。

## (オ) 予測対象メッシュ

予測評価メッシュ数：28,594,240

内訳：居住区メッシュ数 89,357 × 予測対象日数 40日 × 1日当たりの時間分割数 8 = 28,594,240

第2回実験については、当初の設定のとおり予測時間単位を3時間で行った。

また、予測の対象とするメッシュは、第1回実験の「神奈川県内全域」から変更し、「居住区メッシュ」に対象を限定した。

この居住区メッシュは、国土交通省国土数値情報オープンデータで提供される「土地利用区分情報」の「建物用地」メッシュ<sup>(7)</sup>である。

神奈川県における100メートル分割標準メッシュの居住区に区分されるメッシュは89,357メッシュとなり、全体の約35%となる。

メッシュ数としては60%以上の削減となるが、削減される利用区分は森林、河川、海浜、海水域といった本調査研究における予測対象罪種が発生する可能性が低い場所となる。

実際に空き巣のデータで確認すると、実際に発生したメッシュのカバー率は約90%となり、削減率に対して発生場所の未カバー率は低く抑えられた。

メッシュ数を絞り込むことで、得られる効果として考えられるのは

- 予測対象数が減少することによる処理速度の向上。
- 発生メッシュ・未発生メッシュの不均衡（未発生メッシュが多数を占める）という状況の改善。
- 発生可能性の少ないメッシュを対象から除外することによる予測精度の向上。

が考えられる。

予測の発生・未発生の判断<sup>しきい</sup>閾値(基準)については、第1回実験と同じく50%として評価を行った。

## イ 実験結果

### (ア) 予備実験結果

予測対象メッシュ（居住区メッシュ）全体で予測のための学習をさせたところ、サポートベクターマシン（以下「SVM」という）の学習処理が現実的な時間で終了しないという問題が発生した。

SVMは、学習データ量の2乗に比例した学習時間が必要となることから、データ量が増えると急激に学習時間が伸びることが原因と考えられた。

そこで、予備実験として対象罪種を空き巣とし、SVMが現実的な時間内で学習処理が可能なデータ量である1,000件程度に学習データを減らし、4つの機械学習アルゴリズムで予測精度の比較を行った。

予備実験の結果は、以下のとおりである。

◆ 図表2-4 第2回実験（予備実験）結果

アルゴリズム	TP	FP	FN	TN	正解率	適合率	再現率	リスク比	F値
ロジスティック回帰	88	14,602,633	18	11,846,933	44.79%	0.00060%	83.02%	3.97	0.00121%
XGBoost	74	10,152,301	32	16,297,265	61.62%	0.00073%	69.81%	3.71	0.00146%
SVM	90	14,498,814	16	11,950,752	45.18%	0.00062%	84.91%	4.64	0.00124%
ディープラーニング	104	25,259,478	2	1,190,088	4.50%	0.00041%	98.11%	2.45	0.00082%

予測対象日については、絞り込んだ学習データで実際に使用している変数の関係から、予測ができなかった3日分（学習データの大幅な削減のため、予測結果と評価データの不整合が生じたため。）を除外した37日分で評価した。

この結果から、SVMについては現実的な学習データ数で学習するとロジスティック回帰と同程度の予測性能であった。また、一般に大量の学習データ数を必要とすると言われるディープラーニングは、学習データ量が絶対的に不足しているのか、予測が発生方向に大きく振れていた。



XGBoost については、再現率はやや低いものの適合率は高く、少ない学習データ量でも発生の特徴をつかんだ学習ができていると見る事ができた。

以上の予備実験から、SVMは今回の条件下の犯罪発生予測においては

- 実際に運用を考えると、現実的な時間での学習処理が難しい。
- 現実的な処理が可能な件数で比較しても精度はXGBoostに及ばない。

ということが分かった。

そのため、SVMに関してはこの予備実験の結果をもって比較実験の対象から除外することとした。

#### (1) 本実験結果

予備実験の結果を踏まえ、SVM以外の3つの機械学習アルゴリズムを用いて予測実験を行った。これを本実験とする。

モデル	適合率	再現率	精度	F1スコア	学習時間	推論時間	メモリ使用量
XGBoost	0.85	0.75	0.80	0.78	10分	1分	1GB
Random Forest	0.82	0.72	0.78	0.75	15分	1分	1GB
AdaBoost	0.80	0.70	0.75	0.72	20分	1分	1GB

結果は、次の表のとおりである。

◆ 図表2-5 空き巣予測結果 (第2回本実験)

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
ロジスティック回帰	100	17,421,944	16	11,172,179	0.0006%	86.21%	4.01	0.0011%
XGBoost	81	7,783,435	35	20,810,689	0.0010%	69.83%	6.19	0.0021%
ディープラーニング	82	12,624,078	34	15,970,046	0.0006%	70.69%	3.05	0.0013%

◆ 図表2-6 女性・子ども(15歳未満)を対象とした犯罪・前兆事案予測結果 (第2回本実験)

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
ロジスティック回帰	93	15,475,576	4	13,118,567	0.0006%	95.88%	19.71	0.0012%
XGBoost	73	6,844,846	24	21,749,297	0.0011%	75.26%	9.66	0.0021%
ディープラーニング	58	11,064,127	39	17,530,016	0.0005%	59.79%	2.36	0.0010%

◆ 図表2-7 女性・子ども(15歳以上)を対象とした犯罪・前兆事案予測結果 (第2回本実験)

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
ロジスティック回帰	176	12,274,678	34	16,319,352	0.0014%	83.81%	6.88	0.0029%
XGBoost	143	6,760,409	67	21,833,621	0.0021%	68.10%	6.89	0.0042%
ディープラーニング	101	9,939,059	109	18,654,971	0.0010%	48.10%	1.74	0.0020%

◆ 図表2-8 振り込み詐欺前兆電話予測結果 (第2回本実験)

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
ロジスティック回帰	908	14,346,335	597	14,246,400	0.0063%	60.33%	1.51	0.0127%
XGBoost	1,194	4,925,082	311	23,667,653	0.0242%	79.34%	18.45	0.0485%
ディープラーニング	1,218	11,587,338	287	17,005,397	0.0105%	80.93%	6.23	0.0210%

◆ 図表2-9 ひったくり予測結果 (第2回本実験)

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
ロジスティック回帰	24	14,651,587	5	13,942,624	0.0002%	82.76%	4.57	0.0003%
XGBoost	14	6,300,645	15	22,293,566	0.0002%	48.28%	3.30	0.0004%
ディープラーニング	23	14,884,610	6	13,709,601	0.0002%	79.31%	3.53	0.0003%

◆ 図表2-10 自転車盗予測結果 (第2回本実験)

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
ロジスティック回帰	931	17,819,889	34	10,773,386	0.0052%	96.48%	16.55	0.0104%
XGBoost	711	4,491,241	254	24,102,034	0.0158%	73.68%	15.02	0.0316%
ディープラーニング	880	12,858,747	85	15,734,528	0.0068%	91.19%	12.67	0.0137%

第2回の方が第1回よりも問題設定としては難しくなっている（時間の分割数が多い）が、それでも大幅に精度が低下しなかったのは、対象メッシュ数の削減による予測対象数の減少や新たに加えた特徴量が予測結果に貢献しているためと考えられる。

各予測対象罪種について、XGBoost で学習して得られた予測モデルがどの特徴量を重視して予測していたかは以下のとおりである。

◆ 図表2-11 第2回実験特徴量重要度一覧（抜粋）

（上位30位）

	空き巣	振り込め詐欺前兆電話	女性・子ども（15歳未満）を対象とした犯罪・前兆事案	女性・子ども（15歳以上）を対象とした犯罪・前兆事案
1	住宅地面積比率	時間帯	時間帯	時間帯
2	時間帯	住宅地面積比率	住宅地面積比率	気温
3	最寄駐車場までの距離	曜日	最寄小学校までの距離	最寄コンビニまでの距離
4	湿度	ガソリン価格_ハイオク	気温	住宅地面積比率
5	道路面積比率	直近1年での予測対象事象発生件数	最寄公園までの距離	最寄り駅までの距離
6	交通量	気温	最寄中学・高校までの距離	最寄ゴルフ場までの距離
7	直近1年での予測対象事象発生件数	最寄公園までの距離	最寄ゴルフ場までの距離	最寄駐輪場までの距離
8	地区容積率	ガソリン価格_灯油店頭	最寄駐車場までの距離	道路面積比率
9	最寄サービスエリアまでの距離	昼夜間人口比率	交通量	世帯
10	民営借家世帯率	地区容積率	最寄ガソリンスタンドまでの距離	最寄駐車場までの距離
11	最寄交差点までの距離	最寄駐車場までの距離	最寄病院までの距離	最寄中学・高校までの距離
12	ガソリン価格_軽油	月	地区建蔽率	最寄公共施設までの距離
13	気温	風速	最寄商店までの距離	最寄ジャンクションまでの距離
14	一世帯当り延面積	湿度	湿度	直近1年での予測対象事象発生件数

	空き巣	振り込み詐欺前兆電話	女性・子ども（15歳未満）を 対象とした犯罪・前兆事案	女性・子ども（15歳以上）を 対象とした犯罪・前兆事案
15	直近1ヶ月での 予測対象事象発生件数	最寄サービスエリアまでの 距離	最寄ゲームセンターまでの距 離	最寄り駅の乗降客数
16	昼夜間人口比率	最寄駐輪場までの距離	人口増加率	最寄幼稚園までの距離
17	風速	最寄警察署までの距離	最寄ジャンクションまでの距 離	最寄スーパーまでの距離
18	最寄ジャンクションまでの距 離	道路面積比率	最寄ラブホテルまでの距離	最寄空き地までの距離
19	最寄コンビニまでの距離	最寄神社までの距離	男 25～29 歳人口	最寄サービスエリアまでの距離
20	最寄公園までの距離	最寄中学・高校までの距離	地区容積率	最寄ラブホテルまでの距離
21	最寄コインランドリーまでの 距離	最寄小学校までの距離	最寄警察署までの距離	最寄公園までの距離
22	最寄商店までの距離	交通量	最寄公衆浴場までの距離	昼夜間人口比率
23	最寄警察署までの距離	最寄公共施設までの距離	最寄公共施設までの距離	最寄特別支援学校までの距離
24	最寄駐輪場までの距離	最寄文差点までの距離	最寄駐輪場までの距離	最寄協同組合までの距離
25	男 35～39 歳人口	最寄ラブホテルまでの距離	最寄協同組合までの距離	地区容積率
26	最寄古物店までの距離	高齢単身一般世帯	男 45～49 歳人口	最寄深夜飲食店までの距離
27	人口	最寄特別支援学校までの距 離	最寄神社までの距離	駐車場面積比率
28	男 70～74 歳人口	最寄ジャンクションまでの 距離	最寄スーパーまでの距離	国籍
29	性別	人口	最寄空き地までの距離	自営業主 (%)
30	最寄空き地までの距離	最寄自治会までの距離	最寄特別支援学校までの距離	男 35～39 歳人口

全体的に時間帯や気候に関係する項目を重視する傾向があるが、空き巣は住宅地面積比率、地区容積率など、女性・子どもを対象とした犯罪・前兆事案は最寄り駅、コンビニエンスストア、学校までの距離など、既存研究で指摘されている要因を重視しており、罪種に合わせた学習を行い、予測モデルを構築していることが分かる。

一方で、振り込め詐欺前兆電話に関しては、環境的な要因よりも時間帯、曜日などを重視している傾向があった。これは、前兆電話の架電予測であり、だましの電話を掛けるという行為は、電話を受ける側の周囲の環境は何ら影響を与えないためと考えられ、その特徴を適切に学習した結果であると考えられた。

また、ガソリン価格など、現在は学説上、犯罪との関係性が明示されていない特徴量を重視している状況も見られた。

機械学習アルゴリズムの予測性能について比較すると、どの対象罪種に関しても XGBoost が最も高い F 値を示した。

再現率に関しては他のアルゴリズムよりもやや劣る値であったが、リスク比は他のアルゴリズムよりも高い数値を出しており、総合的に考えれば、「発生する」という予測結果が当たりやすいモデルが構築できているという結果となった。

ディープラーニングに関しては、適合率・再現率ともにおおむねロジスティック回帰と同程度の数値に止まる結果となった。

しかしながら予備実験の結果と比較すると精度が相当に向上していることから学習データ数が予測精度に高く影響していることが分かる。

#### (ウ) 時間帯決定方法の検証実験

犯罪発生時刻に幅がある場合の時間決定方法の適否を検証するため、空き巣の予測単位を 24 時間としたモデルでも実験を行った。

◆ 図表 2-12 空き巣予測結果

罪種モデル名	TP	FP	FN	TN	適合率	再現率	リスク比	F値
空き巣 (24時間)	71	1,248,840	45	2,325,324	0.0057%	61.21%	2.94	0.0114%
空き巣 (3時間)	81	7,783,435	35	20,810,689	0.0010%	69.83%	6.19	0.0021%

※使用アルゴリズム:XGBoost

問題設定が異なるため単純に数値で比較することは難しいが、実務上、警戒や検挙活動に活用することを考えると、24時間単位の予測よりも3時間単位の予測の方が利用しやすいと考えられる。

この実験結果の評価から発生時刻に幅がある場合、時間の自至内にランダムに発生時刻を割り振ったとしても3時間幅程度の時間的分類では実用上大きな問題はないと考えられる。また、実務的にも24時間モデルよりメリットが高いものであると考えられる。

### (3) 第3回実験

第2回実験の結果を踏まえ、機械学習アルゴリズム及び特徴量の絞り込み、さらに犯罪の近接反復性を考慮した特徴量の追加等を行い、予測性能の変化を確認することを目的として第3回実験を行った。

実験設定は以下のとおりである。

#### ア 実験設定

##### (ア) 機械学習アルゴリズム

第2回実験の結果から、機械学習アルゴリズムをXGBoost及びディープラーニングに絞り込み、予測精度を比較した。

第2回実験では、ディープラーニングはXGBoostには及ばなかったが、パラメーターを調整することで更なる精度向上の可能性があるディープラーニングと最も精度の良かったXGBoostを用いて実験を行うこととした。

この結果から犯罪予測に相当と考えられる機械学習アルゴリズムを検討する。

##### (イ) 特徴量

第2回実験の特徴量重要度の結果から、人工知能が予測に際して重要度が低かった特徴量を除外し、相関性が高い特徴量は統合及び除外した。

また、予測時の特徴量重要度は高いが、その理由付けが難しい特徴量（ガソリン価格など）については除外することにより特徴量選択を行った。

一方、犯罪間の時空間的な近接性を考慮した特徴量を追加し、予測に影響するかを確認した。

使用した特徴量は別添2「特徴量一覧」の第3回実証実験を参照。

### (ウ) 予測対象罪種

予測対象罪種については、「空き巣」「振り込め詐欺前兆電話」「女性・子ども（15歳未満）を対象とした犯罪・前兆事案」「女性・子ども（15歳以上）を対象とした犯罪・前兆事案」を対象とした。

### (エ) 対象期間、予測対象範囲

対象とした期間・予測対象範囲は、第2回本実験と同一設定とした。

### イ 実験結果

第3回実験の結果は、次のとおりである。

#### ◆ 図表2-13 空き巣予測結果（第3回実験）

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
XGBoost	77	7,518,756	39	21,075,368	0.0010%	66.38%	5.53	0.0020%

※ディープラーニングについては未実施

#### ◆ 図表2-14 女性・子ども（15歳未満）を対象とした犯罪・前兆事案（第3回実験）

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
XGBoost	69	6,713,318	28	21,880,825	0.0010%	71.13%	8.03	0.0021%
ディープラーニング	76	12,764,730	21	15,829,413	0.0006%	78.35%	4.49	0.0012%

#### ◆ 図表2-15 女性・子ども（15歳以上）を対象とした犯罪・前兆事案（第3回実験）

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
XGBoost	149	7,081,995	61	21,512,035	0.0021%	70.95%	7.42	0.0042%
ディープラーニング	184	16,625,407	26	11,968,623	0.0011%	87.62%	5.09	0.0022%

#### ◆ 図表2-16 振り込め詐欺前兆電話（第3回実験）

アルゴリズム	TP	FP	FN	TN	適合率	再現率	リスク比	F値
XGBoost	1,288	6,081,567	217	22,511,168	0.0212%	85.58%	21.97	0.0423%
ディープラーニング	1,126	9,703,903	379	18,888,832	0.0116%	74.82%	5.78	0.0232%



全体的に見ると、第2回実験と比較して適合率・再現率ともに若干ではあるが低下傾向となった。

各予測対象罪種について、XGBoost で学習して得られた予測モデルがどの特徴量を重視して予測していたかは、以下のとおりである。

◆ 図表2-17 第3回実験特徴量重要度一覧(抜粋)

(上位30位)

	空き巣	振り込み詐欺前兆電話	女性・子ども(15歳未満)を対象とした犯罪・前兆事案	女性・子ども(15歳以上)を対象とした犯罪・前兆事案
1	直近1ヶ月以内の最寄予測対象事象発生場所までの距離	時間帯	時間帯	時間帯
2	直近1ヶ月以内の最寄予測対象事象の発生時間帯からの経過時間	直近1ヶ月以内の最寄予測対象事象発生場所までの距離	最寄公園までの距離	直近1ヶ月以内の最寄予測対象事象発生場所までの距離
3	時間帯	直近1ヶ月以内の最寄予測対象事象の発生時間帯からの経過時間	最寄小学校までの距離	気温
4	最寄駐車場までの距離	曜日	直近1ヶ月以内の最寄予測対象事象発生場所までの距離	最寄コンビニまでの距離
5	気温	気温	最寄中学・高校までの距離	昼夜間人口比率
6	最寄交差点までの距離	風速	気温	最寄り駅までの距離
7	湿度	直近1年での予測対象事象発生件数	人口増加率	屋間人口
8	交通量(一般道)	湿度	最寄駐車場までの距離	最寄駐輪場までの距離
9	最寄バス停までの距離	最寄公園までの距離	最寄病院までの距離	道路面積比率
10	最寄警察署までの距離	最寄警察署までの距離	最寄バス停までの距離	国籍外国人比率
11	風速	道路面積比率	屋間人口	最寄駐車場までの距離
12	道路面積比率	昼夜間人口比率	最寄ゲームセンターまでの距離	最寄スーパーまでの距離
13	最寄幼稚園までの距離	最寄神社までの距離	最寄ガソリンスタンドまでの距離	最寄中学・高校までの距離
14	最寄小学校までの距離	高齢単身一般世帯	直近1ヶ月以内の最寄予測対象事象の発生時間帯からの経過時間	最寄幼稚園までの距離
15	最寄神社までの距離	最寄中学・高校までの距離	最寄スーパーまでの距離	最寄公共施設までの距離

	空き巣	振り込め詐欺前兆電話	女性・子ども(15歳未満)を 対象とした犯罪・前兆事案	女性・子ども(15歳以上)を 対象とした犯罪・前兆事案
16	最寄中学・高校までの距離	最寄駐車場までの距離	最寄公共施設までの距離	湿度
17	最寄特別支援学校までの距離	月	交通量(一般道)	交通量(一般道)
18	最寄公園までの距離	最寄特別支援学校までの距離	人口	地区容積率
19	一世帯当り延面積	地区建蔽率	最寄駐輪場までの距離	駐車場面積比率
20	地区建蔽率	最寄空き地までの距離	失業者%	共同住宅1・2階建(%)
21	駐車場面積比率	最寄バス停までの距離	最寄警察署までの距離	最寄り駅の乗降客数
22	民営借家世帯率	最寄小学校までの距離	地区建蔽率	地価
23	最寄協同組合までの距離	最寄幼稚園までの距離	最寄交番・派出所・ 駐在所までの距離	最寄空き地までの距離
24	世帯世帯総数	世帯世帯総数	最寄商店までの距離	最寄特別支援学校までの距離
25	最寄空き地までの距離	最寄公衆浴場までの距離	最寄特別支援学校までの距離	最寄大学までの距離
26	最寄スーパーまでの距離	最寄駐輪場までの距離	最寄古物店までの距離	最寄ガソリンスタンドまでの 距離
27	最寄交番・派出所・ 駐在所までの距離	最寄交差点までの距離	最寄空き地までの距離	風速
28	最寄深夜飲食店までの距離	最寄ホームセンターまでの距 離	最寄公衆浴場までの距離	最寄公園までの距離
29	最寄駐輪場までの距離	人口増加率	最寄ホームセンターまでの距 離	自営業主(%)
30	人口増加率	最寄交番・派出所・ 駐在所までの距離	最寄神社までの距離	最寄商店までの距離

特徴量の重要度を見ると、新たに追加した犯罪間の時空間的な近接性を考慮した特徴量が予測に高く貢献しており、空き巣に関しては重要度の第1位が空間的近接性、第2位が時間的近接性を重視しているという結果となった。

一方、第2回実験で一定の重要度があったものの、理由が説明しづらい特徴量を削減したところ、全体ではやや予測精度は落ちてしまった。

既存研究による知見が得られておらず、現時点では人間が理由を説明することが難しい特徴量であっても、機械学習による学習処理において何らかの相関関係を見だし、予測の根拠としていたことを示す結果となった。

この予測根拠を説明する X A I と呼ばれる技術を適用したが、その点については後述することとする。

機械学習アルゴリズム間の性能比較結果としては、第 2 回実験と同じく、どの対象罪種に関しても XGBoost が高い予測精度を示した。

ディープラーニングはパラメーターの調整を行ったが、XGBoost の精度には追いつけなかった。

### 3 XAIについて

人工知能の目覚ましい進歩を受け、ディープラーニングやアンサンブル学習などの高度な機械学習アルゴリズムが高い予測性能を示す一方で、人工知能を活用する際の問題点の一つとして

- ・人工知能の判断過程の「ブラックボックス」化

が挙げられる。

ディープラーニングをはじめとする昨今の機械学習アルゴリズムは、その学習ロジックが複雑で、予測モデルの解釈も困難であるため

- 答えを導き出した根拠となる特徴量が何だったのか
- それがどのように影響したのか

が人間には理解することができず、判断に至った理由の説明ができないことが大きな問題となってきた。

人工知能を用いて犯罪発生予測を行う場合を考えると、人工知能の予測結果を警察業務に用いるためには、予測の根拠を理解することは不可欠な問題である。

「人工知能がそう予測しているが、その理由はわからない。」という理屈は、地域住民に対しても警察内部に対しても説明責任を果たさなければならない、という点では到底通用しない話だということである。

このような人工知能のブラックボックス化の問題に対して、一定の規制をかける動きも出てきている。

EUでは人工知能が個人に対して下した判断の根拠を求める権利を認める規制（欧州一般データ保護規制（GDPR））が施行された。

日本においても総務省が「AI 開発ガイドライン案」を策定し、その第2条の「透明性の原則」において「開発者は AI システムの入出力の検証可能性及び判断結果の説明可能性に留意すること」と示されるなど、人工知能の説明可能性に関する規制やガイドラインの明文化が進みつつある。

このような人工知能の判断過程のブラックボックス化という課題への対応策として研究が進んでいるのが人工知能の意思決定の根拠を分析する X A I (eXplainable AI : 説明可能な人工知能) と呼ばれる技術である。

X A I 技術の分類としては、

- 学習済の任意のブラックボックス型の人工知能に対してこの X A I を適用し、この人工知能が重視した特徴量や学習データ等を根拠として提示する方法 (Model Induction)
- 人間が理解可能な特徴量や構造を学習させる新しい高性能な機械学習アルゴリズムの開発 (Interpretable Model)

などの方法が提案されており、研究開発が行われている。

本調査研究においては、「SHAP (SHapley Additive exPlanations)」と呼ばれる手法を用いて、どのような特徴量が予測結果にどのように影響を与えたのかを分析することにする。

SHAP とは、Model Induction 型の X A I 技術であり、ゲーム理論 (複数主体の意思決定や行動に関する利得を数学的モデルで表し、分析する理論) に基づき、人工知能の予測値を各特徴量の貢献度の和に分解する手法であり、予測値に対する各特徴量による感度を分析するような方法である。

簡単にいえば、特徴量が少しずつ違うデータを生成して、それらを入力して人工知能に予測をさせた場合に、予測結果がどのように変化するかを観測し、各特徴量の貢献度を算出していくというものである。

現在、SHAP はオープンソースソフトウェアとしても公開されており、誰でも使用することが可能であるが、本調査研究では日立製作所によって一部の処理が高速化されるなど改良版の SHAP を使用することとした。

SHAP を第3回実証実験の予測モデルに適用し、予測根拠を分析した結果のうち、予測に際して重視したものや特徴のあるものをいくつか抽出して添付する。詳細は別添1「SHAP による人工知能の学習傾向分析結果」を参照。

SHAP を適用し、人工知能が予測に際して重視した特徴量を確認した結果、これまでに犯罪との関連性に関する知見がなく、なぜ予測に影響するのか説明が困難な特徴量が影響していることや影響の及び方が予想と逆である特徴量があることが示された。

このような特徴量は、新たな犯罪発生の要因である可能性を示唆するものとも考えられる。しかし、ここで注意しなくてはならないのが人工知能は大量のデータから特徴量間の相関関係を探し出すことは得意とする反面、その因果関係を見いだすことはできないという弱点があるということである。

このため、犯罪予測に関して重要度が高いと人工知能が判断した特徴量が、そのまま犯罪の発生要因であると考えてしまうと誤った判断を生む危険性があることを十分に理解しておかなくてはならない。

人工知能が発生予測に際して重視した特徴量は、「疑似相関」や「逆の因果関係」である可能性を十分に考え、犯罪発生に因果関係のある要因であるのか、単に疑似相関等の関係に過ぎないのかを最終的には人間が判断する必要がある。

その上で、説明性や分かりやすさを重視し、予測精度をある程度犠牲にしても因果関係が確認できる特徴量のみを使用するか、疑似相関の可能性があることを前提としたうえで、予測精度を重視してより多くの特徴量を使用するかは、特徴量の内容を吟味して個別に判断する必要があると考えられる。

いずれにしても、人工知能の予測根拠を示すことが可能な技術は、非常に有益なものであり、今後ますます重要なものとなるであろう。

## 【用語集】

### ○ 特徴量 (Feature)

予測対象にどのような特徴があるかを、人工知能が扱うことが可能なように数値化したものをいう。

面積、気温などの数値そのものに意味のある「数値型」の特徴量と、色や形などのデータに一定の決まりを設けてラベル付けすることで数値化した「カテゴリ型」の特徴量に分けられる。

カテゴリ型は、更にカテゴリ内で大小などの優劣や順序付けが可能な「順序特徴量」と、優劣の関係を持たずに識別子として用いる「名義特徴量」に分けることができる。

### ○ 回帰 (Regression) と分類 (Classification)

過去のデータを用いて将来の値を予測するものを回帰、グループ分けするものを分類という。

犯罪が「発生する」「発生しない」を予測する問題は分類、「何件の犯罪が発生するか」を予測する問題は回帰問題として考えることになる。

分類の中でも「発生する」「発生しない」のような、2つのグループに分けるものを「二値分類」、3つ以上のグループに分けるものを「多値分類」と呼ぶ。

### ○ ハイパーパラメータ

機械学習を行う際、予め設定しておかなければならない変数 (パラメータ) のことをいう。

ハイパーパラメータの設定次第で機械学習の予測精度が変化することがあるので、この調整は重要な要素である。

しかし、機械学習により適切な数値を得られない変数であるため、人間の経験則や試行錯誤で決定していく必要があり、その調整には技術や相当な労力を要する場合がある。



○ 過学習（オーバーフィッティング）

学習データに対する分類能力は高いものの、学習していない未知のデータに対しては正しく分類できない（汎化できていない）状態になることをいう。

「気温」という因子を介して「アイスクリームの販売量」と「水難事故の件数」が疑似的に相関関係を示してしまっているだけであり、実際には何の因果関係もないことから、アイスクリームの販売量を減らしても水難事故が減ることはないので、問題解決にはつながらない。

○ 交差検証（クロスバリデーション）

学習データを分割し、それぞれを順に学習・評価することで過学習していないかを評価する手法のことをいう。

○ L1 正則化

正則化とは、モデルが複雑化することに対してペナルティを課すなどして、過学習を防ぐ手法である。

L1 正則化の場合は、相関の高い説明変数が複数含まれている場合、1つを残して他の変数をゼロとすることで、結果として特徴量選択を行うことになり、汎化能力が向上する。

○ 汎化能力

様々な異なる対象に共通する性質、共通して適用できる法則などを見いだす能力のことをいう。機械学習に関して言えば、学習データに対する分類性能だけでなく、未知の新たなデータに対しても正しく答えを返すことができる能力のことである。

○ アンダーサンプリング

正例と負例のデータ数が不均衡な場合、数の少ないデータに合わせて多数派のデータからランダムにデータを抽出して同数程度に合わせることをいう。

不均衡なデータをそのまま学習させてモデルを作ると、少数派データの分類能力が低くなることが知られているため、アンダーサンプリングを行うことにより少数派のクラスの分類能力が低下することを防ぐ。

## ○ 標準メッシュ (標準地域メッシュ)

標準地域メッシュとは、統計利用のために、緯度・経度に基づいて地域をほぼ同じ大きさの網の目 (メッシュ) に分割したもので、その分割方法については「統計に用いる標準地域メッシュおよび標準地域メッシュ・コード (昭和 48 年行政管理庁告示第 143 号)」により定められている。

総務省統計局などの行政機関が作成しているメッシュ統計に用いる標準地域メッシュは「基準地域メッシュ」「分割地域メッシュ」「統合地域メッシュ」の 3 種類と定め、メッシュの区分とコードの表示方法を定めている。

基準地域メッシュ (3 次メッシュ区画) は、一辺約 1 キロメートル (緯度間隔 30 秒、経度間隔 45 秒) の範囲であり、基準地域メッシュを緯度・経度方向に 2 分割したものを「2 分の 1 地域メッシュ」 (一辺約 500 メートル)、2 分の 1 地域メッシュを更に緯度・経度方向に 2 分割したものを「4 分の 1 地域メッシュ」 (一辺約 250 メートル) として、統計情報が公開されている。

100 メートル分割標準メッシュとは、基準地域メッシュ (3 次メッシュ区画) を緯線方向及び経線方向に 10 等分してできる区域として利用している。

## ○ オープンソースソフトウェア (Open Source Software)

ソースコード (プログラム文) が作者により無償で一般公開されており、一部の例外を除き、商用・非商用の目的を限らずソースコードの利用、修正、拡張、再頒布等が可能となっているソフトウェアをいう。

フリーソフトウェアとは、ソースコードの公開・非公開の違いやプログラムの改変が可能か否かという点で違いがある。

有名なものでは UNIX 系 OS である「Linux」、スマートフォン用 OS である「Android」などがオープンソースソフトウェアとして知られており、多くの個人や企業が利用し、プログラムの改良などを行っている。

数多くの機械学習アルゴリズムや関係するツールなどもオープンソースソフトウェアとして公開されており、本研究においてもこれらのソフトウェアを利用している。

### ○ 疑似相関

実際には2つの事象の間には何の因果関係もないが、双方の間に存在する見えない要因（「交絡因子」「潜在変数」などという。）の介在により、統計的な分析を行うと相関関係が見いだされ、あたかも因果関係があるかのように推測されてしまうというものをいう。

例として「アイスクリームが売れると、水難事故が増える」というものがある。

この2つの事象の間には相関関係があるが、実際にはこの間に「気温」という交絡因子が存在し、これが「アイスクリームの販売量」と「水難事故の件数」のそれぞれに因果関係があるに過ぎない。

### ○ 逆の因果関係

相関関係が存在し、なんらかの因果関係もあるが、原因であると考えられていた事象と結果であると考えられていた事象が実際には逆であるというものをいう。

例として「警察官が増えると、犯罪が増える」というものがある。

警察官が増えたことが原因で結果として犯罪が増えるのではなく、犯罪が増えたことからその対策として警察官の配置を増やしたという原因と結果を取り違えていることから、間違っただけの仮説を導き出してしまった例である。

この仮説に基づいて警察官を減らしてしまうと、普通に考えれば犯罪は増加してしまうと考えられる。

【注釈】

- (1) <https://www.kaggle.com/>
- (2) Python のオープンソース機械学習ライブラリ。様々な機械学習アルゴリズムをサポートしている。 <https://scikit-learn.org/>
- (3) Python による XGBoost をサポートするオープンソースライブラリ。 <https://anaconda.org/anaconda/py-xgboost/>
- (4) Google によるオープンソースソフトウェアライブラリ。ディープラーニングをはじめ、様々な機械学習アルゴリズムをサポートする。 <https://www.tensorflow.org/>
- (5) Tensor-Flow の上部で動作するオープンソースのラッパーライブラリ。 <https://keras.io/ja/>
- (6) 東京大学空間情報科学研究センターが提供するアドレスマッチングサービス。住所が記載されたCSVファイルに対してインターネット経由で緯度・経度の位置情報を付与する。 <http://newspat.csis.u-tokyo.ac.jp/geocode/>
- (7) 国土交通省「国土数値情報 土地利用細分メッシュ」で定義されるもの。 <http://nlftp.milt.go.jp/ksj/jpgis/datalist/KsjTmplt-L03-b.html>

【参考文献】

- 1) Boldot, M., & Borg, A. (2016) 『Evaluating Temporal Analysis Methods Using Residential Burglary Data』、International Journal of Geo-Information 5(9):148
- 2) 西颯人, 樋野公宏『交通事故発生件数のトレンドと季節性の類型化－神奈川県内の車両相互事故を対象として－』、公益社団法人日本都市計画学会都市計画報告集 No.17 pp237-242 2018.8