

# 地域毎の犯罪種別比率を考慮した 畳み込みニューラルネットワークによる犯罪予測 Crime Prediction by Convolutional Neural Network Considering Ratio of Crime Types by Region

向 直人<sup>1)</sup>

Naoto Mukai

## 1 背景

愛知県警察の令和 2 年 犯罪統計書 [1] によると、愛知県における窃盗犯の認知件数は減少傾向にあり、平成 23 年には 91,147 件であったが、令和 2 年には 26,116 件と大幅に減少している。しかしながら、都道府県ごとの自動車盗難件数は、令和 3 年に愛知県がワースト 1 位となっており、警察や地域ボランティアによる防犯活動は今後も欠かすことはできない。平成 28 年に制定された官民データ活用推進基本計画に基づき、都道府県警察は発生年月日、発生時、発生場所などの情報を含む窃盗犯罪の情報を犯罪オープンデータとして公開するようになった。愛知県警察も、犯罪情勢の把握や、地域の防犯力の向上を目的として、平成 30 年から令和 2 年までの犯罪オープンデータを公開している。本研究では、愛知県警察の犯罪オープンデータを利用して、地域の地理的な形状と犯罪傾向の関連性を、機械学習で学習することで、将来の犯罪を予測し、地域の防犯活動に資することを目的とする。

これまでも過去の犯罪データを基に将来の犯罪を予測する研究が報告されている。欧米の先駆的な犯罪予測研究に関する手法やその分類に関しては、大山らが詳しくまとめている [2]。ここでは、日本を対象とした代表的な研究について述べる。野らは、埼玉県さいたま市を対象として、カーネル密度推定と RTM (Risk Terrain Model) を利用して、犯罪発生地点の予測に取り組んだ [3]。説明変数には、駅、基幹道路、駐車場などの 10 種類の環境要因が採用されている。犯罪発生地点を考慮した防犯パトロールを構築し、パトロール担当官を対象としたヒアリング調査の結果、少数による不支持の傾向があるものの、約 80% の肯定的な回答を得たと述べている。また、中川らは、時空間を考慮したカーネル密度推定を利用して、犯罪が頻発する地域 (ホットスポット) を予測する手法を提案した [4]。パトロールに必要なリソースが限られることから、パトロールを実施可能な面積割合を制約条件に用いて、パトロール・カバー率を最大化した。過去の犯罪データを用いた実験では、侵入盗において提案手法が高い有効性を示したと述べている。これまでの犯罪予測手法は、上述の研究でも採用されたカーネル密度推定や RTM が主であるが、近年は深層学習 (ディープラーニング) を利用した研究が報告されている。張らは、犯罪の時間的近接性に着目し、深層学習の一種である LSTM (Long Short Term Memory Network) を利用して、犯罪の時系列データを学習する手法を提案した [5]。この手法では、過去の 6 ヶ月の犯罪件数を説明変数として、当月の犯罪件数を予測する。空き巣、自転車盗、万引きの予測を試みた結果、標準的なニューラルネットワークよりも高い精度で予測が可能であること

1) 椋山女学園大学 文化情報学部

を示した。

上記のいずれの研究においても、過去の犯罪データを基に機械学習のための特徴量を生成し、対象地域の犯罪を予測している。特徴量には、犯罪の発生時や発生場所などの情報に加え、駅や駐車場などの環境要因を用いている。一方で、対象地域の道路・建物の大きさや配置など地理的な条件は特徴量として考慮されていない。我々は、地理的な条件が、特定の犯罪種別の発生に関連していると仮説を立てた。例えば、郊外の閑静な駐車場では、人気が多く「車上ねらい」などの犯罪が発生しやすいと考えられる。また、繁華街では狭隘な道路が複雑に交差していることが多く、見通しが悪いことから「自動販売機ねらい」などの犯罪が発生しやすいと考えられる。そこで、本研究では、画像分類に適した畳み込みニューラルネットワークを利用して、犯罪発生地域の地図画像を学習させることで、対象地域の犯罪種別の比率を予測することを試みる。

本稿の構成は以下である。2 章では、愛知県警が公開している犯罪オープンデータに関して概観する。3 章では、畳み込みニューラルネットワークの学習に用いる地図画像の生成方法に関して述べる。4 章で、提案手法の有効性を検証し、最後に 5 章で本稿をまとめる。

## 2 愛知県の犯罪オープンデータ

本章では、愛知県の犯罪オープンデータの概要を示し、学習用の地図生成に必要な住所から緯度・経度の変換方法について述べる。

### 2.1 概要

愛知県の犯罪オープンデータ [6] には、平成 30 年から令和 2 年までの 3 年間に、愛知県内で発生した 45,456 件の窃盗犯罪に関する情報がまとめられている。犯罪種別 (手口) は、ひったくり、車上ねらい、部品ねらい、自動販売機狙い、自動車盗、オートバイ盗、自転車盗の 7 種類であり、それぞれの犯罪件数を表 1 に示す。愛知県の乗用車保有台数は、2022 年 3 月末において全国第 1 位であり、「車上ねらい」「部品ねらい」「自動車盗」など自動車をターゲットとした犯罪が多いことで知られている。

表 1 愛知県の犯罪種別 (手口) ごとの犯罪件数 (平成 30 年~令和 2 年)

犯罪種別	発生件数
ひったくり	267
車上ねらい	7,070
部品ねらい	5,685
自動販売機ねらい	1,620
自動車盗	2,008
オートバイ盗	1,847
自転車盗	28,059
合計	45,456

## 2.2 ジオコーディング

提案手法では、犯罪発生地点の地理的な要因と、犯罪種別との関連性を明らかにするため、対象地域周辺の地図画像を生成することが必要となる。しかし、犯罪オープンデータには、被害者の個人情報保護の観点から、緯度・経度で表される正確な位置情報は含まれておらず、「都道府県」「市区町村」「町丁目」の文字列情報しか含まれていない。そこで、図 1 に示すように、「都道府県」「市区町村」「町丁目」を連結した住所文字列から、ジオ・コーディング (Mapbox Geocoding API) を利用して、緯度・経度を算出した。しかし、ジオ・コーディングの精度は不十分であり、正確な緯度・経度が算出できない場合がある。このため、算出精度 (relevancy) が 80% 以上であった 39,351 件のデータを学習対象として使用する。

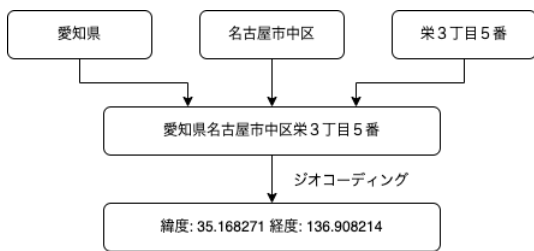


図 1 住所から緯度・経度の変換

## 3 学習用データセットの生成

本章では、前章で述べた愛知県の犯罪オープンデータから、畳み込みニューラルネットワークの学習用データセットを生成する方法について述べる。

### 3.1 犯罪種別の選択

犯罪オープンデータは、7 種類の犯罪種別 (手口) で構成されるが、ここでは問題の簡単化のため、「車上ねらい」「自動販売機ねらい」「オートバイ盗」の 3 種類を対象とする。この 3 種類の犯罪種別は、極端に犯罪件数の多い「自転車盗」を除いて、特定の地域に集中して発生する傾向があり、地理的な要因との関係性が高いと考えた。対象となる 39,351 件の犯罪は、9,520 種類の住所 (緯度・経度) のいずれかで発生している。住所ごとの犯罪発生件数の基本統計量を表 2 に示す。上記の 3 種類の犯罪種別は、最大件数が上位であり、標準偏差も比較的大きいことが読み取れる。

表 2 住所ごとの犯罪発生件数の基本統計量

犯罪種別	平均	標準偏差	最大件数
ひったくり	0.02	0.18	4
車上ねらい	0.62	1.05	19
部品ねらい	0.40	0.77	9
自動販売機ねらい	0.13	0.55	11
自動車盗	0.18	0.48	8
オートバイ盗	0.17	0.52	13

「車上ねらい」「自動販売機ねらい」「オートバイ盗」を合わせた犯罪件数が上位の住所を表 3 に示す。「中区栄」や「中区新栄」など、愛知県名古屋市の繁華街で多くの犯罪が起きていることがわかる。一方で、「自動販売機ねらい」や「オートバイ盗」では、発生件数にバラツキがあり、地理的な要因との関係性が想定される。

表 3 犯罪件数が多い住所の上位 5 件

住所	車上ねらい	自動販売機	オートバイ盗
中区栄 3-5	19	10	13
中区新栄 1-49	15	11	6
中区栄 5-23	11	10	1
中区大須 3-6	11	7	2
中区錦 3-15	10	1	9

対象とする 3 種類の犯罪件数の合計を階級として、各階級の住所数を表すヒストグラムを図 2 に示す。犯罪件数が 1 件だけの住所が最も多く、全体の約 65% (3,318/5,092) を占めていることがわかる。犯罪件数が少ない住所は、地理的な要因との関連性を抽出することが困難であることから、ここでは犯罪件数が 3 件以上であった 780 種類の住所を学習用データセットとして構築する。

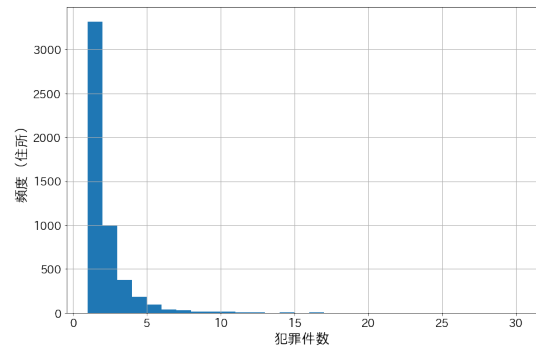


図 2 車上ねらい、自動販売機ねらい、オートバイ盗を合わせた犯罪件数のヒストグラム

### 3.2 k-means 法によるクラスタリング

学習用データセットでは、各住所に対して正解ラベルを設定する必要がある。一つの住所に対し、複数の犯罪が発生しているため、特定の犯罪種別を正解ラベルとすることはできない。そこで、「車上ねらい」「自動販売機ねらい」「オートバイ盗」の犯罪件数を特徴量として、k-means 法によるクラスタリングを適用し、導出されたクラスタを正解ラベルとする。特徴量は範囲変換を用いて最小 0、最大 1 となるように正規化する。学習に用いるサンプル数が極端に少なくなることを防ぐために、クラスタ数を  $k=3$  に固定した結果、重心が図 3、図 4、図 5 で表されるクラスタが算出された (便宜上、クラスタ A、B、C と表す)。クラスタ A は、141 件の住所が該当し、「オートバイ盗」が最も多く、次いで「車上ねらい」が多く発生している。クラスタ B は、489 件の住所が該当し、「車上ねらい」の発生件数が突出している。クラスタ C は、150 件の住所が該当し、「自動販売機ねらい」が最も多く、次いで「車上ねらい」が多く発生している。

### 3.3 地図画像の生成

住所に対応する緯度・経度から、Mapbox Static Tiles API を利用し、地図画像を生成する。生成する地図画像のサイズは  $256 \times 256$  であり、ズームレベルを 17 に設定したことから、地図画像の一辺は約 234 メートルである。表 4 に示す 5 種類の色で、地図画像の領域は色分けされている。駐車場は点データであるため、同一半径の円で表すことにした。また、駐車場は Mapbox で参照可

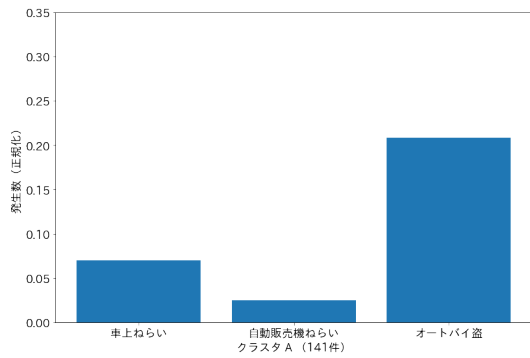


図 3 クラスタ A の重心

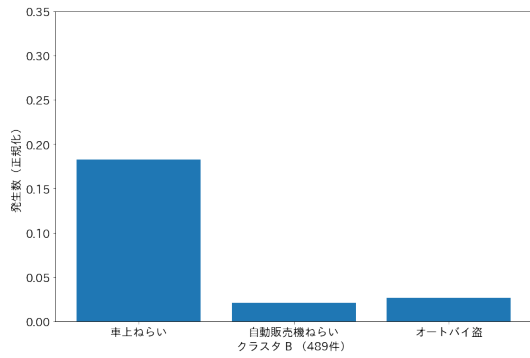


図 4 クラスタ B の重心

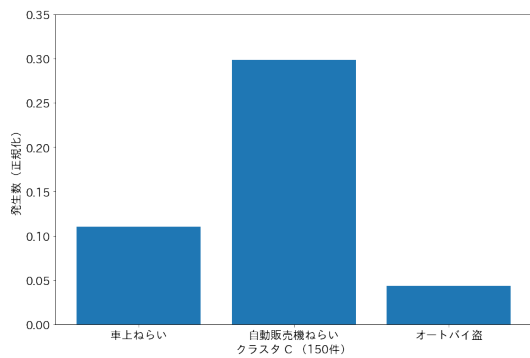


図 5 クラスタ C の重心

表 4 地図画像の色分け

色	領域
白	道路領域
黒	背景領域
緑	建物領域
赤	駅領域
青	駐車場 (点データ)

能なデータから抽出しているが、全ての駐車場が含まれているわけではないことに注意されたい。

図 6 は、クラスタ A に属する「中区栄 2-14」の地図画像である。左は参考のためのオープンストリートマップの地図画像であり、右が本手法で生成した学習に用いる地図画像である。地図中に 3 箇所の駐車場の存在が確認でき、クラスタ A の特徴であった「オートバイ盗」の発生に関連があると予想される。図 7 は、クラスタ B に属する「尾張旭市東本地ヶ原町 2」の地図画像である。先の地図画像とは異なり、郊外の住宅地であることが確認できる。クラスタ B の特徴であった「車上ねらい」は、自宅駐車場に停められた自動車を狙った犯行である可能性が考えられる。図 8 は、クラスタ C に属する「中区栄 3-5」の地図画像である。当該地図は名古屋の繁華街の中心であり、高層ビルや大きな公園が地図の大部分を占めている。クラスタ C の特徴であった「自動販売機ねらい」のターゲットとなる自動販売機が多く設置されていると思われる。上記で述べたような地理的な特徴量を、畳み込みニューラルネットワークに学習させることで、任意の地図に対し、起こりうる犯罪種別の傾向が推定できる可能性がある。



図 6 クラスタ A に属する中区栄 2-14 (35.166021, 136.902116) の地図

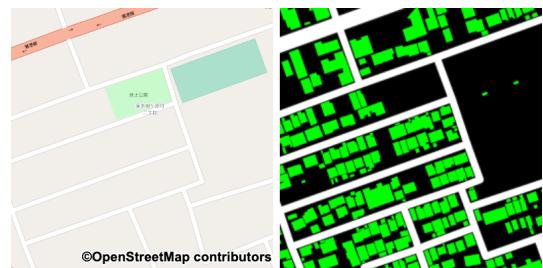


図 7 クラスタ B に属する尾張旭市東本地ヶ原町 2 (35.198288, 137.043143) の地図



図 8 クラスタ C に属する中区栄 3-5 (35.168271, 136.908214) の地図

#### 4 実験

提案手法の有効性を確認するため、畳み込みニューラルネットワークによる予測精度を評価した。表 5 に示すように、790 件のサンプルの 9 割を訓練用データ、残りの 1 割を評価用データとして用いた。畳み込みニューラルネットワークとして、G.Huang 氏らが提案した DenseNet[7] を採用した。DenseNet は勾配消失問題の回避に優れ、効率的な学習を実現したモデルである。地図画像は  $64 \times 64$  に縮小し、データ拡張のためランダムに上下左右に反転させて学習させる。評価関数は交差エントロピー損失、最適化手法は Adam を採用し、 $N = 32$  のミニバッチ学習でモデルを最適化させた。

表 5 訓練用データと評価用データのサンプル数

クラス	訓練用データ	評価用データ
クラス A	127	14
クラス B	440	49
クラス C	135	15
合計	702	78

図 9 は損失、図 10 は正解率の推移を示している。学習のスタート時には、訓練用データに合わせて評価用データの損失が減少し、正解率が上昇していることが確認できる。しかし、訓練用データの損失が大きく減少するに連れ、評価用データの損失は上昇し、正解率も減少してしまう。これは、オーバーフィッティングが原因であると考えられ、同一のクラスタにおいて、共通の特徴量が少ないことを示唆している。評価用データの正解率は最大で 67.9% (53/78) であり、クラス B のサンプル数の比が 62.8% (49/78) であることを考慮すると、5.1% の改善しか確認できなかった。

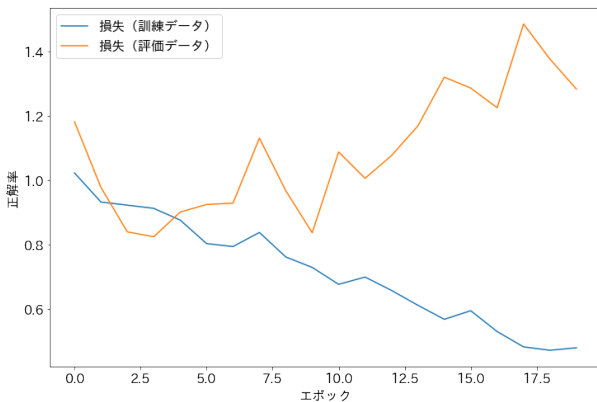


図 9 損失の推移

精度改善が十分に得られなかった原因を考察する。第 1 にサンプル数が不十分であったことが考えられる。本稿では、3 件以上の犯罪発生を条件としたことから、対象となる住所が限定されてしまい、地図画像の特徴を満遍なく学習できなかった可能性がある。愛知県だけの犯罪オープンデータを対象とするのではなく、他の都道府県警察の犯罪オープンデータをマージして、学習に用いるなどの工夫が必要だと考える。第 2 に地図画像の表現方法の改善が考えられる。本稿で生成した地図画像では、道路や建物などの基本的な構造に、環境要因である駅と駐車場を合わせて表現した。しかし、他の研究で

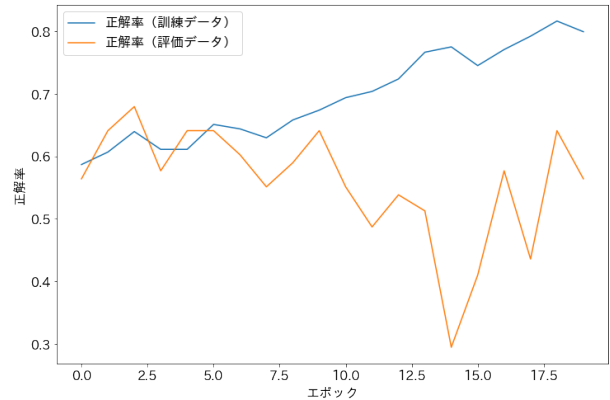


図 10 正解率の推移

は、これらに加え、コンビニエンスストア、金融機関、低層建物、高層マンションなどの環境要因を考慮している。そこで、建物の高さなどの情報を、地図の色で表現するなどの工夫が考えられる。

#### 5 まとめ

本稿では、愛知県警察が公開している犯罪オープンデータを利用して、「車上ねらい」「自動販売機ねらい」「オートバイ盗」の 3 種類の犯罪種別の比率で表されたクラスタを、対象の地図画像を学習させた畳み込みニューラルネットワークで予測する手法を提案した。地図画像は、道路や建物などの基本的な構造に加え、環境要因である駅と駐車場を表現しており、これらの地理的な特徴が、特定の犯罪種別を誘発していると考えた。実験の結果、最大で 5.1% の正解率の改善が確認できたが、オーバーフィッティングにより正解率は減少し、安定した分類精度を実現することができなかった。同一のクラスタにおいて、共通の特徴量が少ないことが原因の一つであると考えられるが、サンプル数や環境要因の不足も影響していると考えられる。今後は、犯罪件数の多い東京などの犯罪オープンデータをマージして、サンプル数を増やして評価する必要がある。また、駅や駐車場に加え、犯罪と高い関連性があるとされる環境要因を考慮することで、分類精度の改善を図りたい。

#### 参考文献

- [1] 愛知県警察令和 2 年犯罪統計書. <https://www.pref.aichi.jp/police/anzen/hassei/keiji-s/images/Aichi-HanzaiTokei2020.pdf>.
- [2] 大山智也, 雨宮護, 島田貴仁, 中谷友樹. 地理的犯罪予測研究の潮流. *G I 理論と応用*, Vol. 25, No. 1, pp. 33–43, 2017.
- [3] 野貴泰, 糸井川栄一. 犯罪多発地点の予測に基づく防犯パトロール経路に関する提案. *地域安全学会論文集*, Vol. 31, No. 0, pp. 195–204, 2017.
- [4] 中川淳子, 西村祥治, 宮野博義. 犯罪発生履歴データの機械学習による時空間カーネル密度推定型犯罪予測の最適化. *情報処理学会論文誌*, Vol. 61, No. 7, pp. 1255–1266, jul 2020.
- [5] 張曉齊, 米澤剛, 吉田大介. オープンデータと lstm を用いた犯罪発生の予測及び時間的近接性における考察. *情報学*, Vol. 16, No. 1, pp. 56–70, jun 2019.
- [6] 愛知県警察愛知県警察犯罪オープンデータサイト. <https://www.pref.aichi.jp/police/anzen/toukei/opendata/seian-s/crimeopendata.html>.
- [7] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.