

# 犯罪発生要因の可視化と不均衡なデータに 対処した予測モデルの精度向上

**Visualizing Crime Factors and Improving the Accuracy of  
Predictive Models Dealing with Imbalanced Data**

山本 藤也 (Touya Yamamoto)  
u220067@st.pu-toyama.ac.jp

富山県立大学 情報システム工学科 情報基盤工学講座

9:35-9:50, Tuesday, February 14, 2023  
N212, Toyama Prefectural University

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

# 1.1 本研究の背景

2/17

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

- 犯罪は、完全にランダムで発生するわけではなく、その場所の地理的な特徴など、何らかの要素が影響していると考えられている。
- 過去に発生した犯罪のデータや、犯罪の要因となり得るデータを用いて、いつ、どこで犯罪が発生するかを予測する研究が、欧米を中心に盛んである。



図1 犯罪が発生する3要素

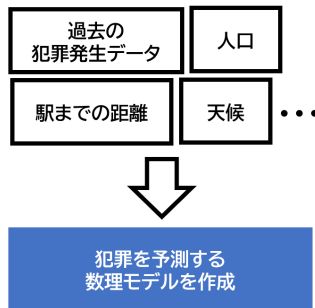


図2 地理的犯罪予測の概要

## 1.2 本研究の目的

3/17

- 欧米と比較して犯罪が低頻度なわが国においても、適切なアプローチを行うことで、時空間的解像度および精度を高く予測することを目指す。
- 地図上に要因を可視化し、根本的な防止へ知見を得ることを支援する。



図3 わが国を取り巻く現状



図4 本研究の目的

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

## 2.1 地理的犯罪予測の概要

4/17

- 地理的犯罪予測は、1980年代から欧米を中心に研究されはじめ、現在に至るまで、さまざまな手法が開発されている。

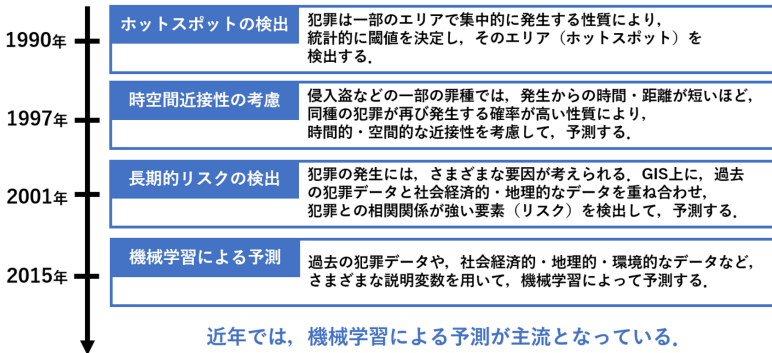


図5 地理的犯罪予測の大まかな動向

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

## 2.2 長期的リスクと短期的リスク

5/17

- 犯罪が発生しやすいところ・ときには、たとえば、レジャー施設の近くや、大規模なイベントが開催されているときなど、さまざまな要因が考えられている。

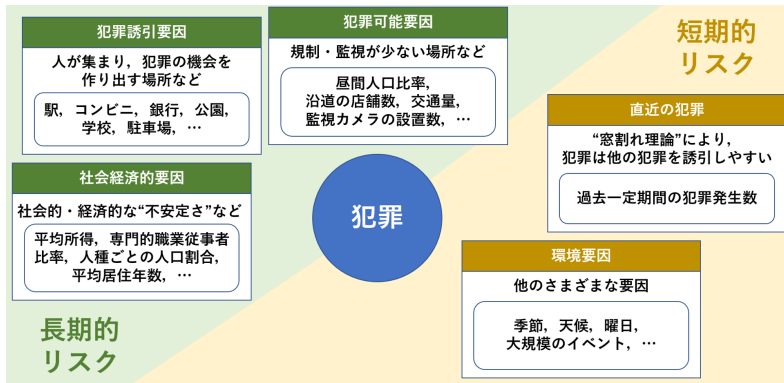


図6 犯罪が発生する「長期的リスク」と「短期的リスク」

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

## 2.3 地理的犯罪予測と GIS

6/17

- GIS の発展により、犯罪の発生地点を地図上に可視化し、ホットスポットを検出できるようになったことが、地理的犯罪予測の源流といえる。

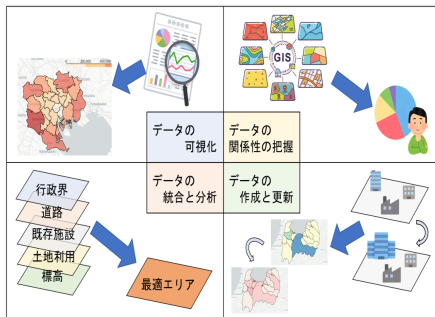


図7 GISの特徴

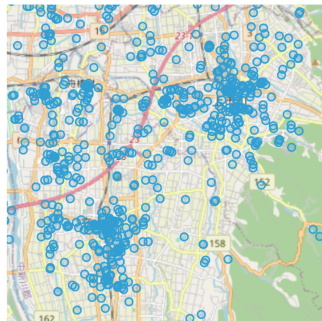


図8 犯罪発生地点を地図上に可視化したようす

- はじめに
- 犯罪とリスク
- 機械学習による犯罪予測と要因分析
- 提案手法
- 数値実験並びに考察
- おわりに

# 3.1 不均衡なデータに対する機械学習

7/17

- 機械学習で分類問題を扱うときは、一般に、各クラスのサンプル数の差が大きいほど、精度は小さくなる傾向がある。

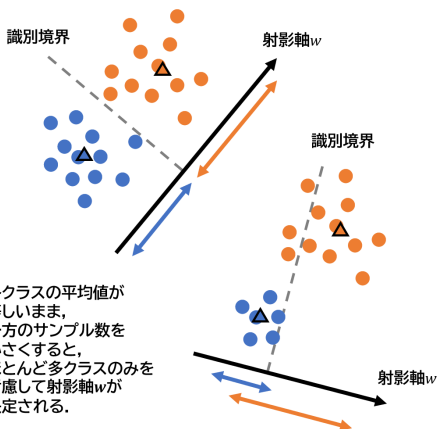


図9 不均衡なデータのための線形判別分析

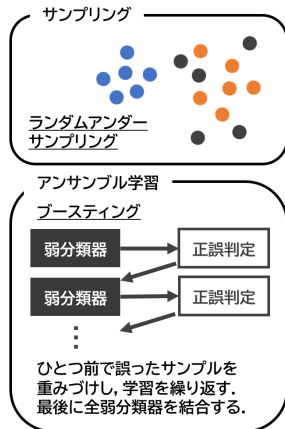


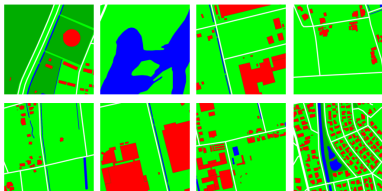
図10 サンプリングとアンサンブル学習

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

## 3.2 サイバー空間からの多様なデータの取得

8/17

- 空き地の面積や、金融機関までの距離など、地理的な特徴を表すデータを、地図画像や地図情報サイトから取得する。



↓ 画像処理で、構造化データへ

メッシュコード	建物の面積比率	空き地の面積比率	道路
0001	0.324	0.521	
0002	0.761	0.109	
0003	0.231	0.653	
0004	0.113	0.721	
0005	0.008	0.897	

図11 地図画像にもとづく説明変数の抽出



↓ スクレイピング、ジオコーディング

コンビニエンスストア	金融機関	レジャー施設	
[137.203, 36.603]	[136.310, 36.321]	[137.433, 36.692]	[137.4
[137.214, 36.651]	[137.002, 36.621]	[137.562, 36.832]	[137.5
[137.195, 36.599]	[137.322, 36.691]	[137.203, 36.293]	[137.2
[137.201, 36.709]	[136.893, 36.761]	[137.543, 36.654]	[137.5
[137.195, 36.541]	[137.312, 36.782]	[137.143, 36.932]	[137.1

図12 地図情報サイトからのスクレイピング



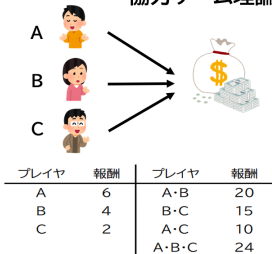
## 3.3 予測モデルを解釈する手法

9/17

- 複雑なモデルであっても、その予測値を出力した理由を解釈できる手法として、Shapley Additive Explanations (SHAP) がある。

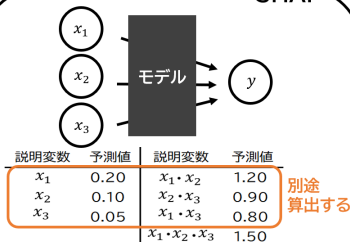
1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

### 協力ゲーム理論



プレイヤーの報酬が、単純な加算ではないとき、各プレイヤーの“貢献度”を表す指標として、**Shapley値**がある。

### SHAP



Shapley値を機械学習のモデルに応用したものが**SHAP**であり、各説明変数の目的変数への“貢献度”を**SHAP値**とよぶ。

$$\phi_j = \frac{1}{|J|!} \sum_{S \subseteq J \setminus \{j\}} (|S|! (|J| - |S| - 1)!) (v(S \cup \{j\}) - v(S))$$

$\phi_j$ : プレイヤjのShapley値

$J$ : プレイヤの集合  $|S|$ : 集合sの要素数

$S$ : 集合jからプレイヤjを除いた集合

$v(\cdot)$ : 報酬を表す関数

図13 SHAPの概要

## 4.1 多様な要因を考慮したデータセットの作成

10/17

- 長期的リスクをまとめた2次元テーブルに、短期的リスクを加えた3次元テーブルを作成する。

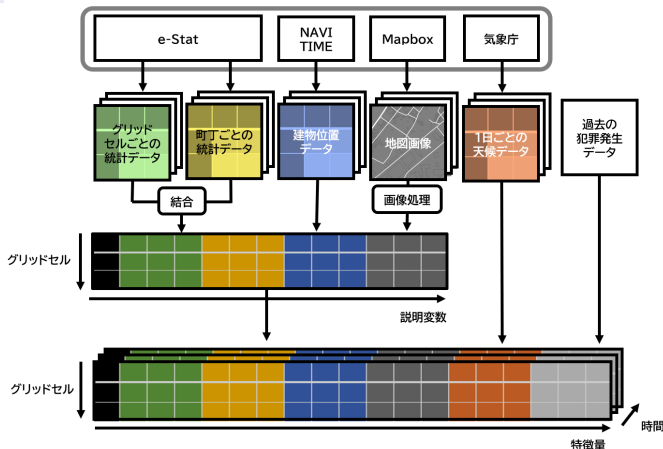


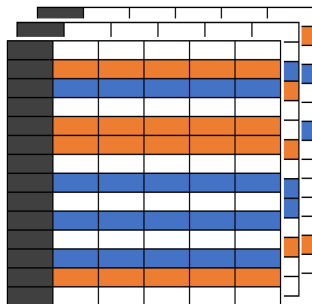
図14 データセットを作成するまでの流れ


1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに


## 4.2 不均衡なデータに対処した予測モデルの構築

11/17

- 1日ごとにランダムアンダーサンプリングを行う。目的変数に対して有意に寄与する説明変数を選択し、アンサンブル学習を行う XGBoost で予測モデルを構築する。



 犯罪が発生したサンプル

 ランダムアンダーサンプリングで選択された  
犯罪が発生していないサンプル

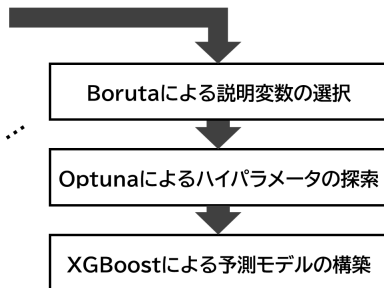


図15 予測モデルを構築するまでの流れ

## 4.3 犯罪発生要因の可視化

12/17

- 長期的リスクについて、予測モデルに対して SHAP を適用し、SHAP 値のテーブルを作成する．それを、地図上に可視化する．

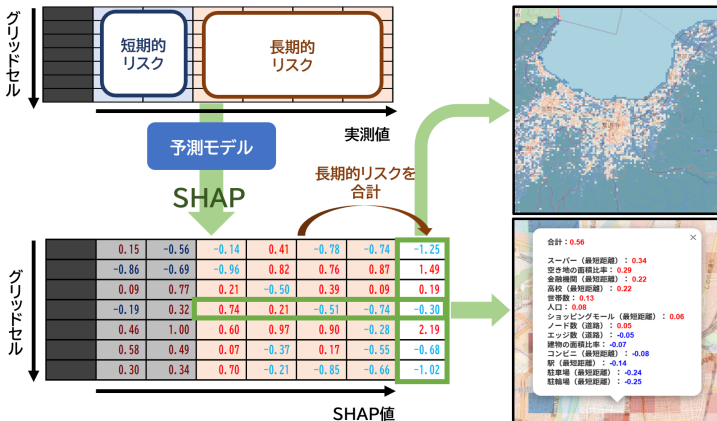


図16 犯罪発生要因を可視化するまでの流れ

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

## 5.1 数値実験の概要 (1)

13/17

- 富山県で 10 年間に発生した路上犯罪 25,120 件を学習用データとし、1 か月間に発生した 157 件を予測する。
- 長期的リスク 59 個、短期的リスク 8 個を説明変数とする。

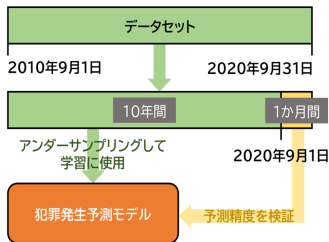


図17 数値実験の流れ

表1 データセットに含まれる発生件数

カテゴリ	発生件数
自転車盗難	13121
声かけ・つきまとい	6760
車上荒らし	5211
タイヤ盗難	3968
部品ねらい	690
わいせつ	401
ひったくり	59

		予測値	
		0	1
実測値	0	TN	FP
	1	FN	TP

$$\text{再現率} = \frac{TP}{TP+FN} \quad \text{適合率} = \frac{TP}{TP+FP}$$

$$F_1\text{スコア} = \frac{2TP}{2TP+FP+FN}$$

図18 評価指数

## 5.1 数値実験の概要 (2)

14/17

表2 説明変数一覧

人口	コンビニエンスストア(立地数)	デパート(立地数)	道路の面積比率
18歳未満人口割合	コンビニエンスストア(最短距離)	デパート(最短距離)	建物の面積比率
65歳以上人口割合	駅(立地数)	警察署/交番(立地数)	空き地の面積比率
外国人人口割合	駅(最短距離)	警察署/交番(最短距離)	水の面積比率
世帯数	駐車場(立地数)	小学校(立地数)	ノード数(道路)
単身世帯割合	駐車場(最短距離)	小学校(最短距離)	エッジ数(道路)
核家族世帯割合	駐輪場(立地数)	中学校(立地数)	密度(道路)
正規労働者割合	駐輪場(最短距離)	中学校(最短距離)	平均次数(道路)
非正規労働者割合	金融機関(立地数)	高校(立地数)	平均気温
最終学歴が中学以下の人口割合	金融機関(最短距離)	高校(最短距離)	日照時間
最終学歴が高校の人口割合	旅館(立地数)	大学(立地数)	降水量
最終学歴が大学以上の割合	旅館(最短距離)	大学(最短距離)	降雪量
居住年数5年未満の人口割合	ホテル(立地数)	保育園/幼稚園(立地数)	過去7日間の犯罪発生件数
居住年数20年以上の人口割合	ホテル(最短距離)	保育園/幼稚園(最短距離)	過去1か月間の犯罪発生件数
一戸建て世帯割合	スーパーマーケット(立地数)	レジャー施設(立地数)	曜日(ダミー変数)
アパート・低中層マンション世帯割合	スーパーマーケット(最短距離)	レジャー施設(最短距離)	休日(ダミー変数)
高層マンション割合	ショッピングモール(立地数)	長期的リスク	短期的リスク
	ショッピングモール(最短距離)		



Borutaによって選択された説明変数

## 5.2 実験結果と考察 (1)

15/17

表3 検証用データによる  
予測精度

		予測値	
		0	1
実測値	0	90946	2677
	1	107	50

(犯罪が発生したセルを1, していないセルを0とする)

適合率	0.018
再現率	0.318
$F_1$ スコア	0.035

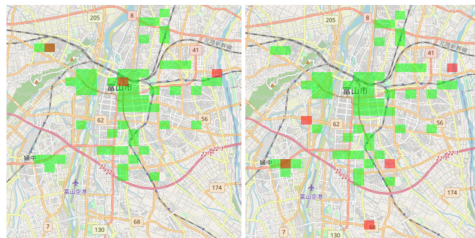


図18 2020年9月1日(左)と2日(右)の予測結果

表4 予測値への寄与が大きい上位5つの説明変数

説明変数	SHAP値の絶対値の平均
過去1か月間の犯罪発生件数	0.34
世帯数	0.22
スーパー (最短距離)	0.20
駐車場 (最短距離)	0.16
空き地の面積比率	0.15

- 予測精度は、必ずしも大きいといえない結果となった。
- 地図画像から抽出した、空き地の面積比率、建物の面積比率、道路ネットワークのエッジ数が、予測値に寄与していた。

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

## 5.2 実験結果と考察 (2)

16/17

1. はじめに
2. 犯罪とリスク
3. 機械学習による犯罪予測と要因分析
4. 提案手法
5. 数値実験並びに考察
6. おわりに

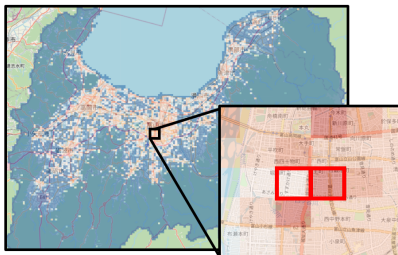


図19 検証するグリッドセル

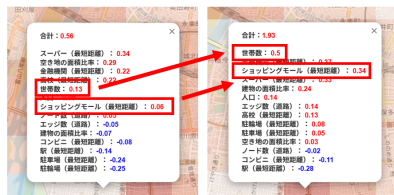


図20 グリッドセルごとに要因を可視化した結果

- となり合うセルで、一方のセルが、他方の約3.45倍のリスクがあると予想された。特に、世帯数が多いこと、ショッピングモールまでが近いことが寄与していると予想された。  
→ 単純に犯罪発生件数を分布したときでは得られないであろう知見を得られた。



## 6. おわりに

17/17

### まとめ

本研究では、犯罪の発生が低頻度なわが国においても、適切なアプローチを行うことによって、時空間的解像度や精度を高く犯罪を予測することを試みた。また、予測モデルを解釈することによって、地図上に犯罪が発生する要因を可視化した。

- 予測精度は、必ずしも大きいといえない結果となった。
- 予測モデルを解釈した結果を地図上に可視化することで、単に犯罪発生件数を分布したときには得られないであろう知見を得られた。

### 今後の課題

- 予測精度の向上
  - 短期的リスクとなる説明変数の追加や、重みづけを行う。
  - 教師なし学習の異常検知問題として扱う。
- 予測モデルを解釈した結果には、当然ながら誤差が含まれるため、予測モデルを介さず、要因を可視化する手法の検討。