

# OUTLIER DETECTION TECHNIQUE TO EXTRACT CANDIDATES OF FRAUD FROM MEDICAL INSURANCE CLAIMS (医療保険請求データからの詐欺候補 者検出のための外れ値検出法)

物理統計学分野  
河盛亮介

# 目次

- 研究背景
- 先行研究
- データ
- 提案手法
- 評価指標
- 結果
- まとめと今後の方針

# 研究背景

- 米国の一 年間の支出のうち少なくとも700億ドル-2340億ドルが詐欺により消失[1]
- 高齢化社会により医療費は増大
- 医療の高度化により医療行為は多様化

→ 質の高い医療を提供するために詐欺発見・阻止システムが必要

- 実データを用いた詐欺発見のシステムは発展途上
- 詐欺行為者のデータも未だ少ない
- 専門家による調査はコストが大きい

→ 外れ値検出により詐欺候補者を抽出し、専門家による調査を支援

[1][https://www.pcmanet.org/wp-content/uploads/2016/08/pr-dated-05-09-13-whitepaper\\_oct10.pdf](https://www.pcmanet.org/wp-content/uploads/2016/08/pr-dated-05-09-13-whitepaper_oct10.pdf)

# 研究背景

[2]<https://www.nhcaa.org/resources/health-care-anti-fraud-resources/consumer-info-action.aspx>

## 詐欺とは

An intentional deception or misrepresentation that the individual or entity makes knowing that the misrepresentation could result in some unauthorized benefit to the individual, or the entity or to some other party.[2]

## 詐欺事例

- ニュージャージー州にて，在宅医療医師が実際より長い時間のサービスを申請し500,000ドル以上を不正受給



# 先行研究

## 教師あり学習

- ニューラルネット[3]
- 決定木[4]
- 重み付き異常度による分類[5]
- Random Forest, ロジスティック回帰, SVM[6]

[3] P.A.Ortega,C.J.FigueroaandG.A.Ruz,"A medical claim fraud/abuse detection system based on data mining: a case study in Chile." In Proceedings of International Conference on Data Mining, Las Vegas, Nevada, USA, 2006.

[4] F.Bonchi,F.Giannotti,G.MainettoandD.Pedreschi,"A classification-based methodology for planning auditing strategies in fraud detection," In Proceedings of SIGKDD99, pp. 175184, 1999.

[5] H. Shin, H. Park, J. Lee and W. C. Jhee, "A scoring model to detect abusive billing patterns in health insurance claims," Expert Systems with Applications, vol 39(8), pp. 74417450, 2012.

[6] R. A. Bauder and T. M. Khoshgoftaar. "The detection of medicare fraud using machine learning methods with excluded provider labels ", in FLAIRS Conference, 2018, pp. 404409.

## 教師なし学習

- ページランク[7]
- ドメイン知識を用いたもの[8]
- Isolation Forest, Local Outlier Factorなど[9]

[7] Jiwon Seo, Mendelevitch O. "Identifying frauds and anomalies in Medicare-B dataset", Conf Proc IEEE Eng Med Biol Soc. 2017.

[8] G. Van Capelleveen, M. Poel, R. M. Mueller, D. Thornton and J. Van Hillegersberg. "Outlier detection in healthcare fraud: A case study in the Medicaid dental domain," International Journal Of Accounting Information Systems, 2016, 21, pp. 18-31.

[9] Richard Bauder, Raquel da Rosa, and Taghi Khoshgoftaar. Identifying medicare provider fraud with unsupervised machine learning. In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pages 285–292. IEEE, 2018.



教師なし学習の詐欺検出に注目

# 先行研究

## 米国医療保険 Medicare

- 障害者・65歳以上の高齢者を対象とした医療保険
- 2017年のコストは、7,060億ドル
- パートA - パートDが存在
- 53,403,309人の被保険者

## Medicare Provider Utilization and Payment Data

- 医療費請求のデータ
- 氏名・住所・郵便番号・医療行為ごとの年間医療費請求額・回数・患者数・実際に支払われた額など
- 2015年は9,497,892行

	npi	zip	hcpcs_code	average_submitted_chrg_amt
1	1003000126	21502	99217	328.0000
2	1003000126	21502	99219	614.0000
3	1003000126	21502	99221	333.2881

# 先行研究 データ

## Richardらの手法[9]

[9] Richard Bauder, Raquel da Rosa, and Taghi Khoshgoftaar. Identifying medicare provider fraud with unsupervised machine learning. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 285–292. IEEE, 2018.

### ■ Medicare Provider Utilization and Payment Data

- 医療費請求のデータ
- 氏名・住所・郵便番号・医療行為ごとの年間医療費請求額・回数・患者数・実際に支払われた額など
- 2015年・9,497,892行

### ■ List of Excluded Individuals/Entities (LEIE)

- 登録除外された医療行為者のデータ
- 1977-2018年まで70,491件
- 除外日・除外理由などを記録

NPI	EXCLDATE	EXCLTYPE	BUSNAME	LASTNAME	FIRSTNAME
1 0000000000	19880830	1128a1	14 LAWRENCE AVE PHARMACY	NA	NA
2 0000000000	19970620	1128b7	143 MEDICAL EQUIPMENT CO	NA	NA
3 1922348218	20180419	1128a1	184TH STREET PHARMACY CORP	NA	NA

# 先行研究

## Richardらの手法[9]

[9] Richard Bauder, Raquel da Rosa, and Taghi Khoshgoftaar. Identifying medicare provider fraud with unsupervised machine learning. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 285–292. IEEE, 2018.

### 手法

- Medicare Provider Utilization and Payment Dataの50%のデータ
- 特徴量: 全医療行為の年間治療回数・患者数・日毎の患者数/日毎の治療回数・年間医療費請求額・年間医療費の標準偏差・最小値・最大値・中央値・合計値・平均値, 男女のカテゴリ, 専門科のカテゴリ
- LOF・KNN・AE・IF・URFで外れ値検出
- ラベルとしてLEIEを使用し評価

### 評価

TPR・TNR・ROC曲線・AUC

# 先行研究 特徴ベクトル

## Richardらの手法[9]

[9] Richard Bauder, Raquel da Rosa, and Taghi Khoshgoftaar. Identifying medicare provider fraud with unsupervised machine learning. In 2018 IEEE International Conference on Information Reuse and Integration (IRI), pages 285–292. IEEE, 2018.

医療提供者の全医療行為から統計量を計算し、特徴ベクトルを作成

NPI	hcpcs_code	Provider_gender	Provider_type	average_medicare_submitted_amount
101111	91312	M	Internal Medicine	10
101111	91313	M	Internal Medicine	10
101111	91314	M	Internal Medicine	25
101112	91312	F	Ophthalmology	50



医療提供者ごとの特徴ベクトルへ

NPI	provider_gender	provider_type	average_medicare_submitted_amount_mean	average_medicare_submitted_amount_sum
101111	M	Internal Medicine	15	45
101112	F	Ophthalmology	5	50



専門科による治療の違い・地域による多様性が考慮されにくい

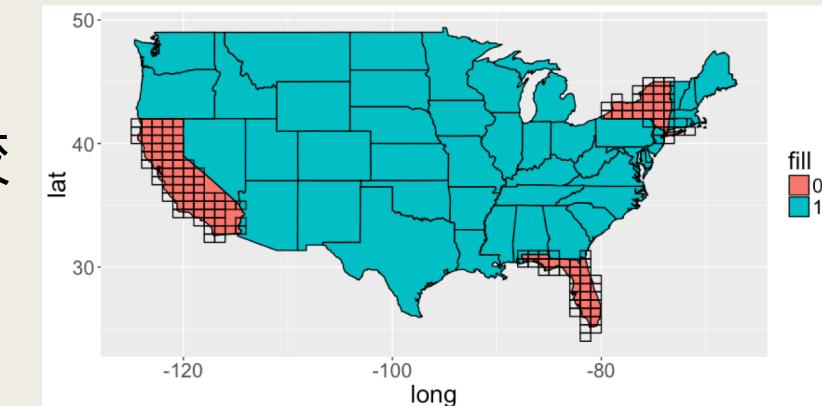
# 提案手法

## 前処理

- Medicare Provider Utilization and Payment Data
- New York・Californiaのデータ
- 1,111,686行, 123299の医療提供者, 46の登録除外者

- データを  $\begin{cases} \text{専門科} \\ \text{専門科+地域} \end{cases}$  で分割, 精度を比較

- 各医療手続きの請求額を特徴量
- One support vector machine・Local Outlier Factorを外れ値検出に使用



## 評価

TPR・TNR・ROC曲線・AUC

# 提案手法 データ分割

## ■ Medicare Provider Utilization and Payment Data

- 医療費請求のデータ
- 氏名・住所・郵便番号・医療行為ごとの年間医療費請求額・回数・患者数・実際に支払われた額など
- 2015年・9,497,892行

## ■ Zipcode Data

- Zipcode に対する緯度経度
- 43,689行

	npi	zip	hcpcs_code	average_submitted_chrg_amt
1	1003000126	21502	99217	328.0000
2	1003000126	21502	99219	614.0000
3	1003000126	21502	99221	333.2881

Medicare Provider Utilization and Payment Data

	zip	longitude	latitude
1	00210	-71.01320	43.00590
2	00211	-71.01320	43.00590
3	00212	-71.01320	43.00590

Zipcode Data

# 提案手法 データ結合



# 提案手法 特徴ベクトル

各医療手続きの請求額 (HCPCS)

医療提供者の各種医療行為請求額の特徴ベクトルを作成

npi	hcpcs_code	average_medicare_submitted_amount	line_svc_cnt
101111	91312	10	3
101111	91313	10	4
101111	91314	25	2
101112	91312	50	2



医療提供者ごとの特徴ベクトルへ 対数+正規化

NPI	91312	91313	99314
1011111	30	40	50
1011112	0	0	100

# One-Class SVM[10]

[10] Schölkopf B., Platt J., Shawe-Taylor J., Smola A.J., and Williamson R.C.2001. Estimating the support of a high-dimensional distribution. Neural Computation,13(7): 1443-1471.

$(\mathbf{w} \cdot \Phi(\mathbf{x})) - \rho \leq 0$ となる点を外れ値とする.  
損失関数は $\max\{0, \rho - (\mathbf{w} \cdot \Phi(\mathbf{x}_i))\}$ とする.

最小化する関数は

$$\frac{1}{vl} \sum_i \max\{0, \rho - (\mathbf{w} \cdot \Phi(\mathbf{x}_i))\} - \rho + \frac{1}{2} \|\mathbf{w}\|^2$$

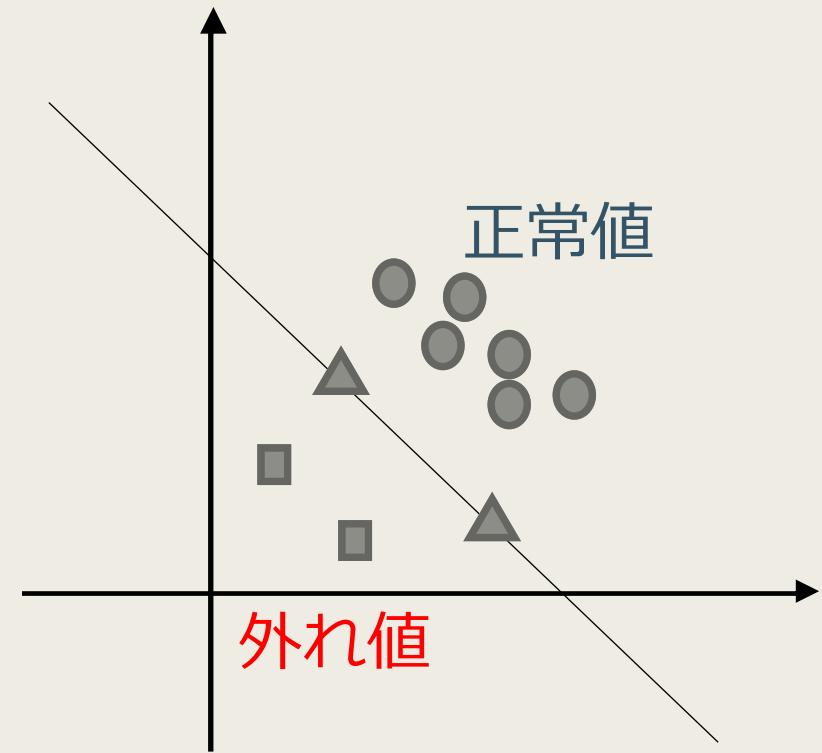
損失                    切片                    正則化



One-Class SVMの主問題

$$\min_{\mathbf{w} \in F, \xi \in \mathbb{R}^l, \rho \in \mathbb{R}} \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{vl} \sum_i \xi_i - \rho$$

$$\text{subject to } (\mathbf{w} \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \xi_i \geq 0.$$



# Local Outlier Factor[11]

[11] BREUNIG, M. M., KRIEGEL, H.-P., NG, R. T., AND SANDER, J. 1999. Optics-of: Identifying local outliers. In Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery. Springer-Verlag, 262–270.

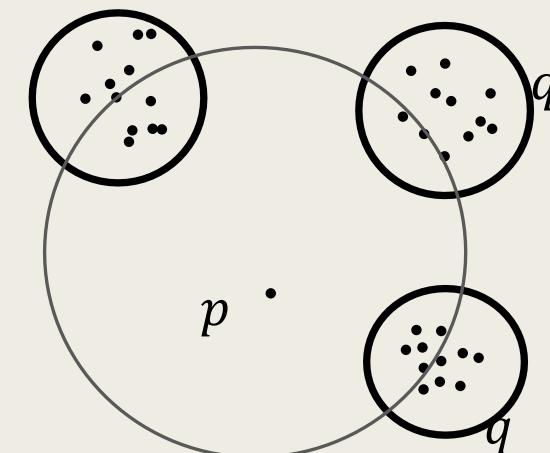
$LOF$

$$LOF_{MinPts}(p) = \frac{\sum_{q \in N_{MinPts}(p)} \frac{lr d_{MinPts}(q)}{lr d_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

点 $p$ の近傍点の周辺の密度  
注目する点 $p$ 周辺の密度

$LOF(p)$ はデータ点 $p$ の異常度

$p \in \mathbb{R}^l$ : 注目するデータ点  
 $q \in \mathbb{R}^l$ :  $p$ の近傍のデータ点

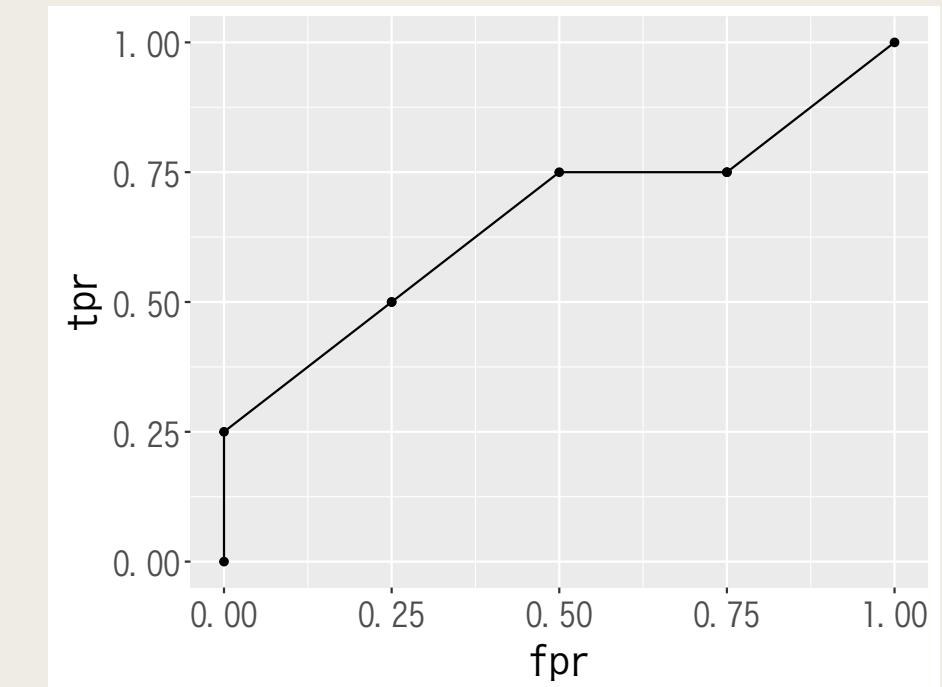


高次元空間

# 評価指標

- $TPR = \frac{TP}{TP+FN}$  +のうち+と予測した割合
- $FPR = \frac{FP}{FP+TN}$  -のうち+と予測した割合
- ROC曲線: 横軸にFPRを縦軸にTPRをプロットした曲線
- AUC: ROC曲線の下部面積

異常度	ラベル
10.3	異常
9.2	異常
8.6	正常
7.5	異常
6.2	正常
5.1	正常
5.0	正常
4.9	異常



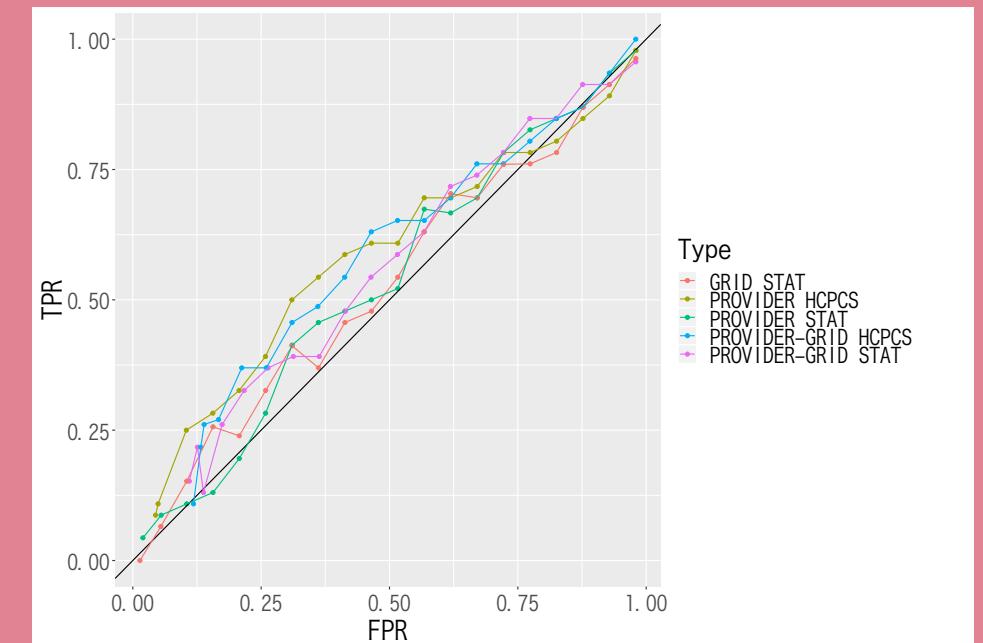
# 結果

## SVMによる外れ値検出

Data-Split	Feature	AUC
NONE	STAT	NA
PROVIDER	STAT	0.513
PROVIDER	HCPCS	0.562
PROVIDER-GRID	STAT	0.535
PROVIDER-GRID	HCPCS	0.561
GRID	STAT	0.509
GRID	HCPCS	NA

データ分割法・特徴量ごとのAUC

- ランダムよりAUC高
- HCPCS: 各治療ごとの請求額の特徴ベクトルがAUC高
- STAT: 各治療の要約統計の場合はPROVIDER-MESH分割が有効



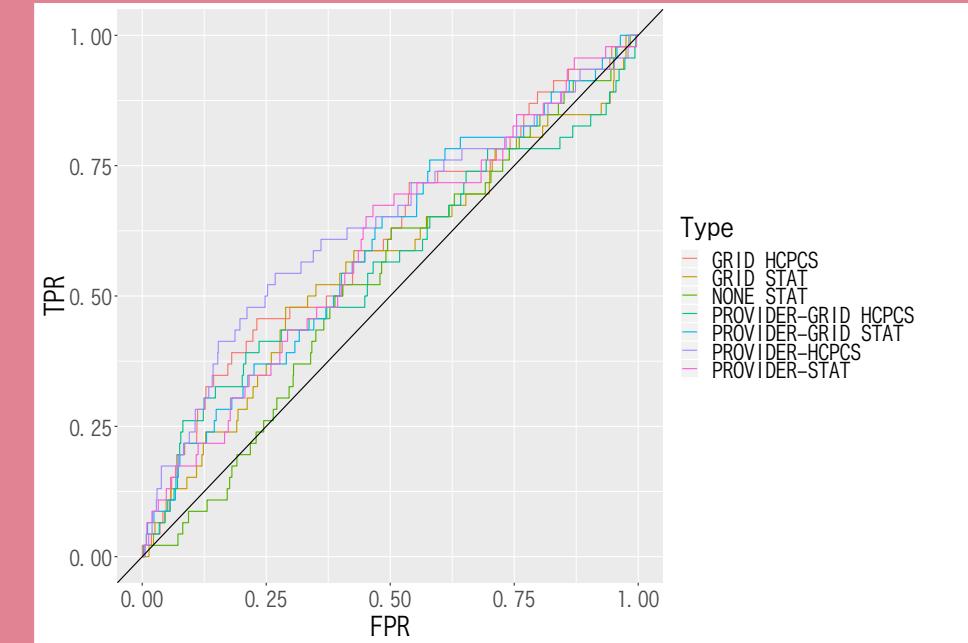
データ分割法・特徴量ごとのROC曲線

# 結果

## LOFによる外れ値検出

Data-Split	Feature	AUC
NONE	STAT	0.529
PROVIDER	STAT	0.591
PROVIDER	HCPCS	0.634*
PROVIDER-GRID	STAT	0.593
PROVIDER-GRID	HCPCS	0.559
GRID	STAT	0.560
GRID	HCPCS	0.605

データ分割法・特徴量ごとのAUC



データ分割法・特徴量ごとのROC曲線

- 先行研究よりAUC高 (ただし、データセット異)
- どのデータ分割も有効
- HCPCS: 各治療ごとの請求額の特徴ベクトルがAUC高
- STAT: 治療の要約統計の場合はPROVIDER-MESH分割が有効

# まとめと今後の方針

## まとめ

- 外れ値と詐欺行為者の一致を実験
- 特徴量・分割の有効性を検証
- 無作為抽出・先行研究より精度良

## 今後の課題

- 専門家にインタビュー
- 他の外れ値検出手法との比較
- どの程度のFPR・TPRが許容されるか