

健診データを用いた 生活習慣病の発症予測

恒川充¹ 岡夏樹¹ 荒木雅弘¹
新谷元司² 吉川昌孝³

1 京都工芸繊維大学

2 SGホールディングスグループ
健康保険組合

3 日本システム技術株式会社



背景

- 昨今のネット通販の普及により、宅配件数が増加
→ドライバーの**健康状態**の管理が重要
- 発症予測の機運の高まり
ex)心筋梗塞や脳梗塞の発症確率を予測

[Yatsuya et .al 2016]

目的

医療データを機械学習に利用



生活習慣病の発症を予測



事故リスクの軽減、医療費の抑制

データの概要と予測対象

- SGホールディングスグループ健康保険組合の医療データを使用

利用したデータの概要

	年代	年齢層	人数	枚数
レセプトデータ	1996~2017	15~74	156,145	961,906
健診データ	2006~2018	15~74	108,581	1,617,078

- 予測対象として定義した重症化病名：
糖尿病，狭心症，心筋梗塞，心筋症，心房細動，
心室細動，くも膜下出血，脳内出血，脳梗塞

用意されているデータ

発症のタイミングを判断

➤ レセプトデータ (★)

ー 医療報酬の明細書

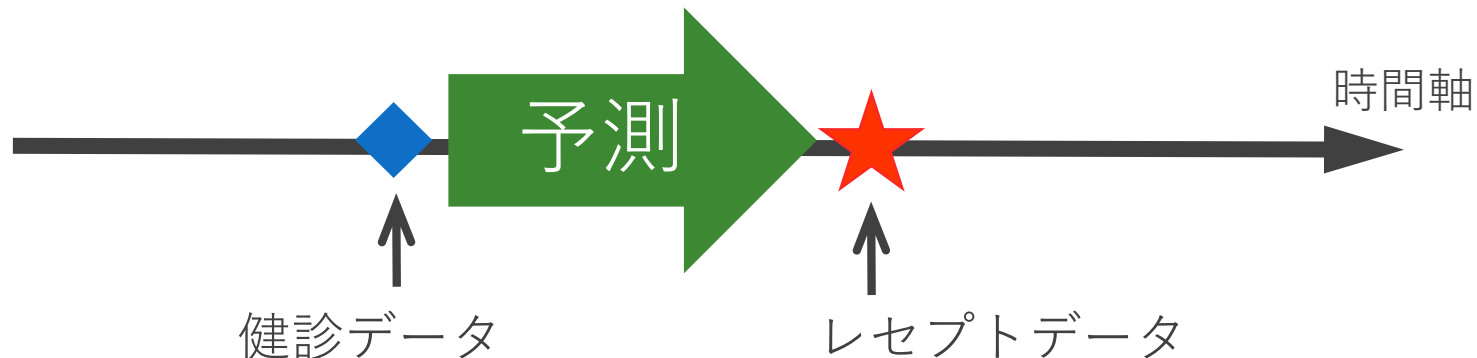
診療年月, 診断病名, 処方された薬 etc.

発症するか否かを識別

➤ 健診データ (◆)

ー 健康診断の結果。

身長, 体重, 血圧, 赤血球数 &
問診表の回答結果 & 判定結果(6段階)



データの特徴

- 重症化病名が初出であると断定できない
 - ・ 中途採用者が存在
 - ・ 健康保険組合に加入している時期のデータしかない
- 正例データと負例データの認定手順が煩雑
- データの偏り
 - ・ 健康な人のデータ >> 病気の人データ
(負例) (正例)
 - ・ 重症化病名の割合は全体の4.5% (2017年)

病気診断データの選定

- ~~レセプトデータ上で病名を見つける
⇒ 正例に用いるデータとする~~

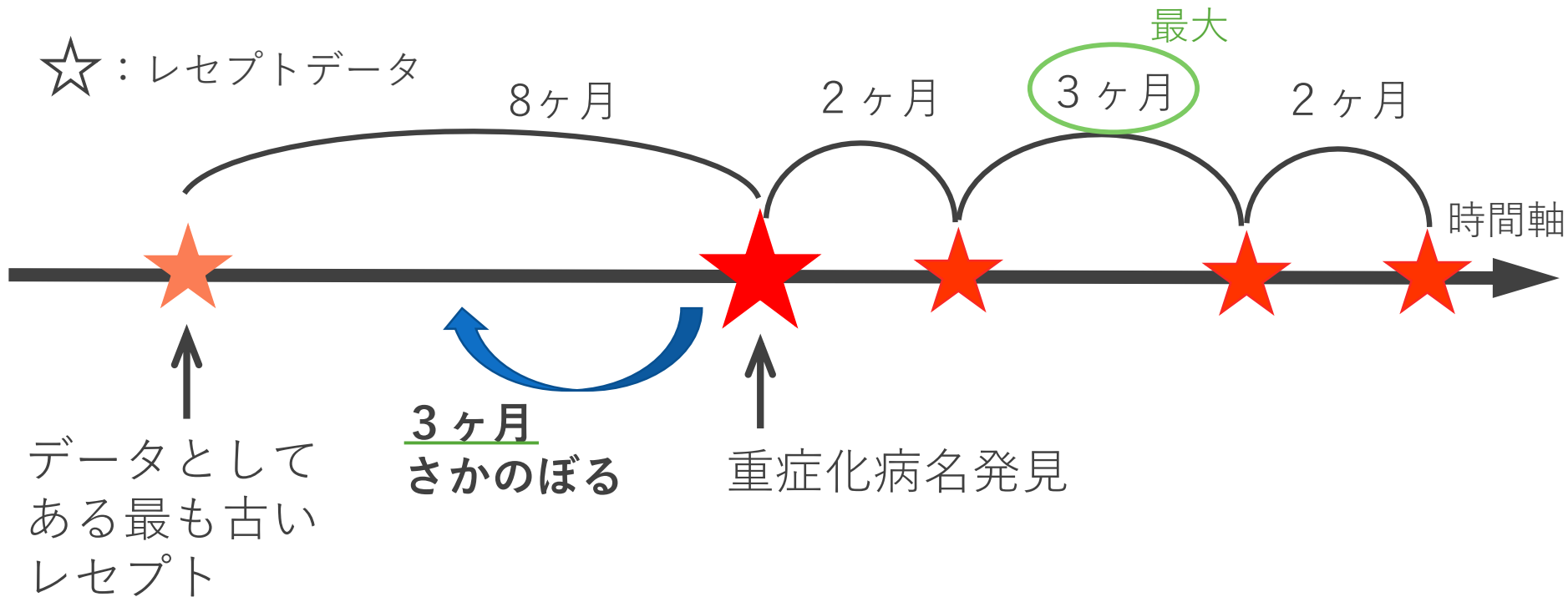
なぜなら…

検査をするために便宜的に病名をつける

- ・ 「疑い病名」 は取り除く
- ・ 薬と病名の対応を確認する



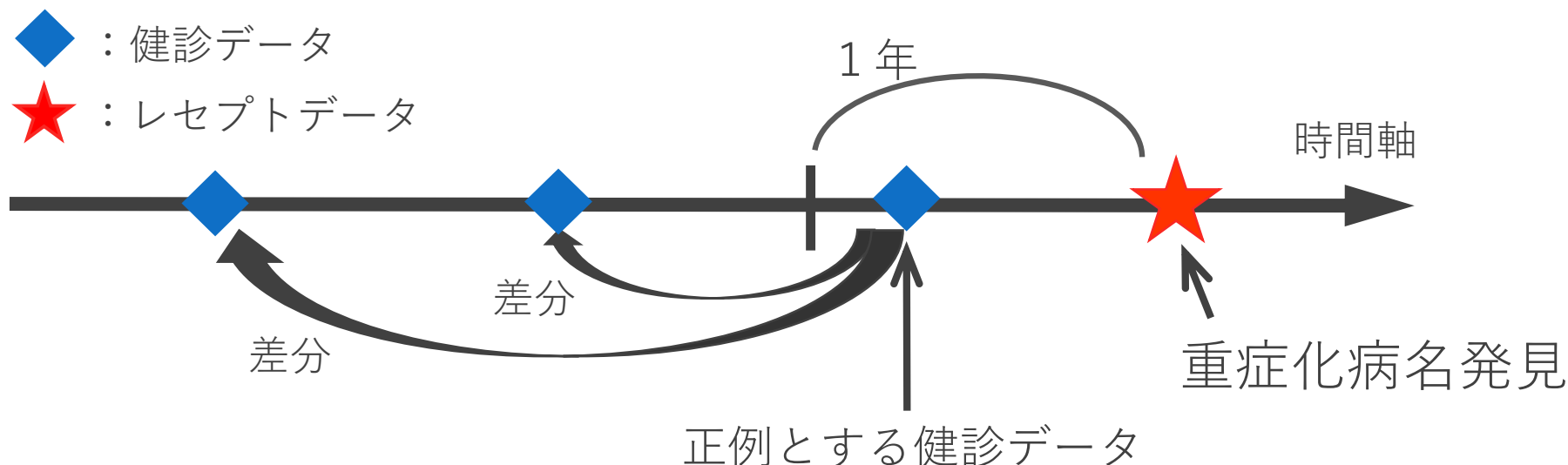
病名を初出とみなす条件



◎推定した通院間隔より長期間過去にレセプトデータが存在する
→重症病名を持って保険に加入してきたわけではない

病気の人(正例)データの作成

- 問題設定：「1年以内に重症化するか否か」



- データの形：
$$\frac{\text{健診データそのもの}}{60\text{次元}} + \frac{\text{一つ目との差分}}{36\text{次元}} + \frac{\text{二つ目との差分}}{36\text{次元}}$$

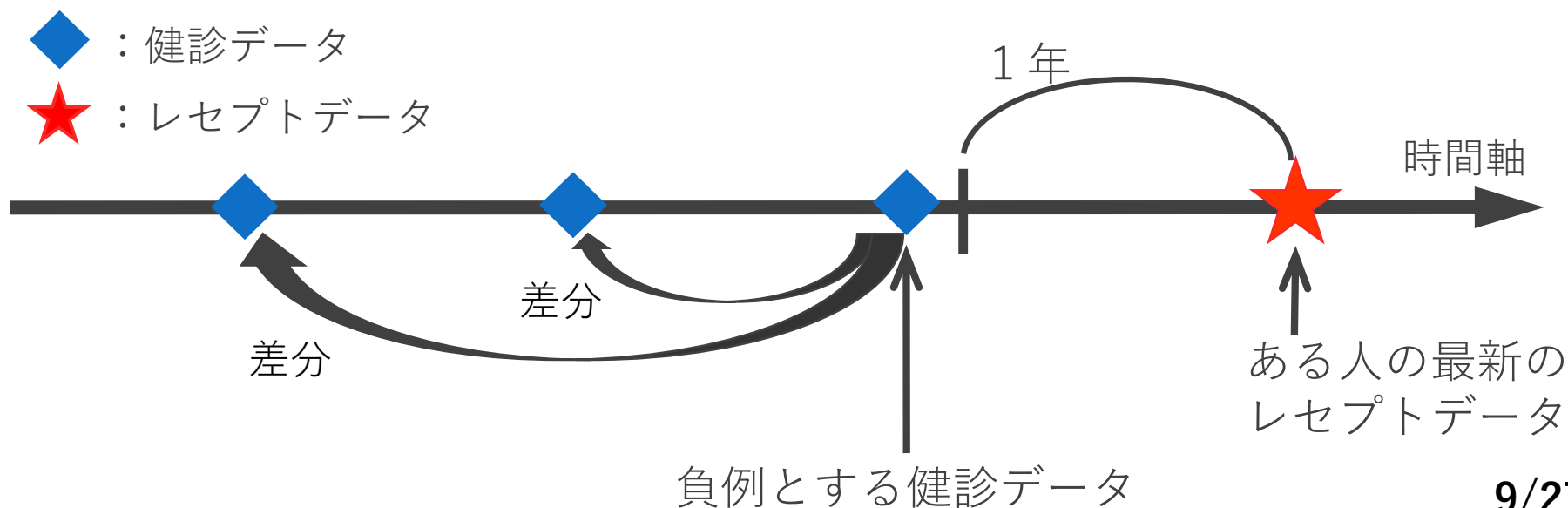
- 病気を発症する際には、何らかの項目に変化がある
⇒ 差分に注目

健康な人(負例)データの作成

➤ 負例データ

- ・ 重症化病名の対象である病名が一度でもついた人間を除外
- ・ 同様に健診データ3つを使って差分を計算して特徴量に追加

➤ 問題設定(1年以内に重症化するか否か) を保証するための工夫



整形後のデータの概要

➤ データサイズ

正例データ	1255
負例データ	37664

不均衡データ：アンダーサンプリング＋バギング

➤ 弱識別器の数：500

➤ 特徴量：132

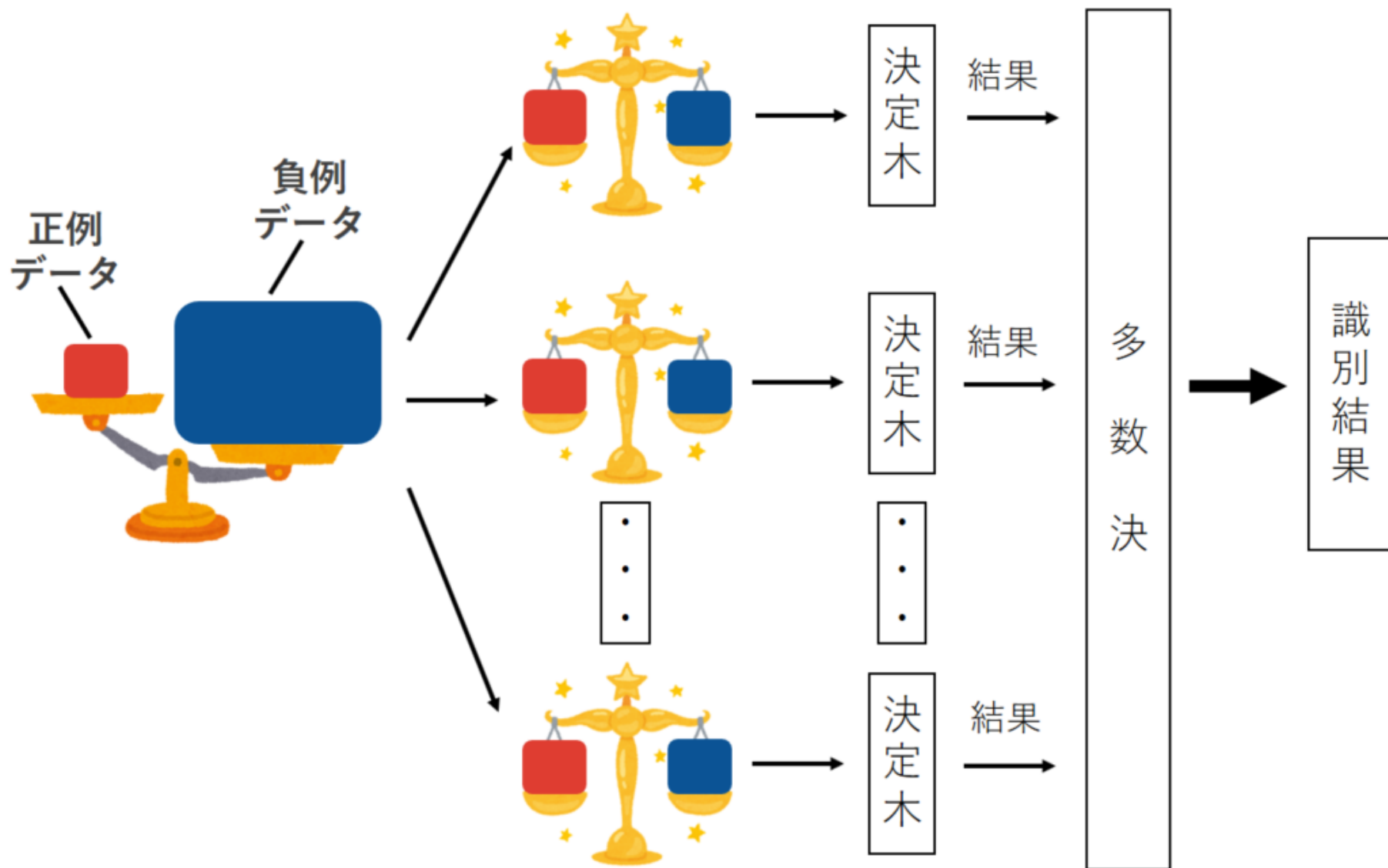
➤ 欠損値：中央値で補完

➤ 標準化処理はしない

∵ 決定木→スケールに影響されないアルゴリズム

➤ 層化10分割クロスバリデーション

アンダーサンプリング + バギング



整形後のデータの概要

➤ データサイズ

正例データ	1255
負例データ	37664

不均衡データ：アンダーサンプリング＋バギング

➤ 弱識別器の数：500

➤ 特徴量：132

➤ 欠損値：中央値で補完

➤ 標準化処理はしない

∵ 決定木→スケールに影響されないアルゴリズム

➤ 層化10分割クロスバリデーション

結果

➤ Confusion Matrix

		予測されたクラス	
		正例	負例
実際の クラス	正例	1118	137
	負例	2306	35358

※ recall :

本当の正例のうち、正例と予測できたものはどれくらいか

precision :

予測した正例のうち、
本当の正例はどれくらいか

* 正例の**recall : 0.89**

正例のprecision : 0.33

➤ 比較手法として…

- ・ 日本人間ドック学会の判定区分表 (13項目, 3段階に分類)
 - － 閾値を設定→OR条件
- ・ 13項目だけを使って提案手法で識別

結果

➤ Confusion Matrix

		予測されたクラス	
		正例	負例
実際の クラス	正例	1118	137
	負例	2306	35358

※ recall :

本当の正例のうち、正例と予測できたものはどれくらいか

precision :

予測した正例のうち、
本当の正例はどれくらいか

* 正例の**recall : 0.89**

➤ 比較手法として…

- ・ 日本人間ドック学会の判定区分表 (13項目, 3段階に分類)
 - － 閾値を設定→OR条件
- ・ 13項目だけを使って提案手法で識別

結果

➤ Confusion Matrix

		予測されたクラス	
		正例	負例
実際の クラス	正例	1118	137
	負例	2306	35358

※ recall :

本当の正例のうち、正例と予測できたものはどれくらいか

precision :

予測した正例のうち、
本当の正例はどれくらいか

* 正例の**recall : 0.89**

正例のprecision : 0.33

➤ 比較手法として…

- ・ 日本人間ドック学会の判定区分表 (13項目, 3段階に分類)
 - － 閾値を設定→OR条件
- ・ 13項目だけを使って提案手法で識別

結果

➤ Confusion Matrix

		予測されたクラス	
		正例	負例
実際の クラス	正例	1118	137
	負例	2306	35358

※ recall :

本当の正例のうち、正例と予測できたものはどれくらいか

precision :

予測した正例のうち、
本当の正例はどれくらいか

*

正例のprecision : 0.33

➤ 比較手法として…

- ・ 日本人間ドック学会の判定区分表 (13項目, 3段階に分類)
 - － 閾値を設定→OR条件
- ・ 13項目だけを使って提案手法で識別

結果

➤ Confusion Matrix

		予測されたクラス	
		正例	負例
実際の クラス	正例	1118	137
	負例	2306	35358

※ recall :

本当の正例のうち、正例と予測できたものはどれくらいか

precision :

予測した正例のうち、
本当の正例はどれくらいか

* 正例の**recall : 0.89**

正例のprecision : 0.33

要医療、要経過観察、軽度異常

➤ 比較手法として…

- ・ 日本人間ドック学会の判定区分表 (13項目, 3段階に分類)
ー 閾値を設定→OR条件
- ・ 13項目だけを使って提案手法で識別

ベースライン手法との比較

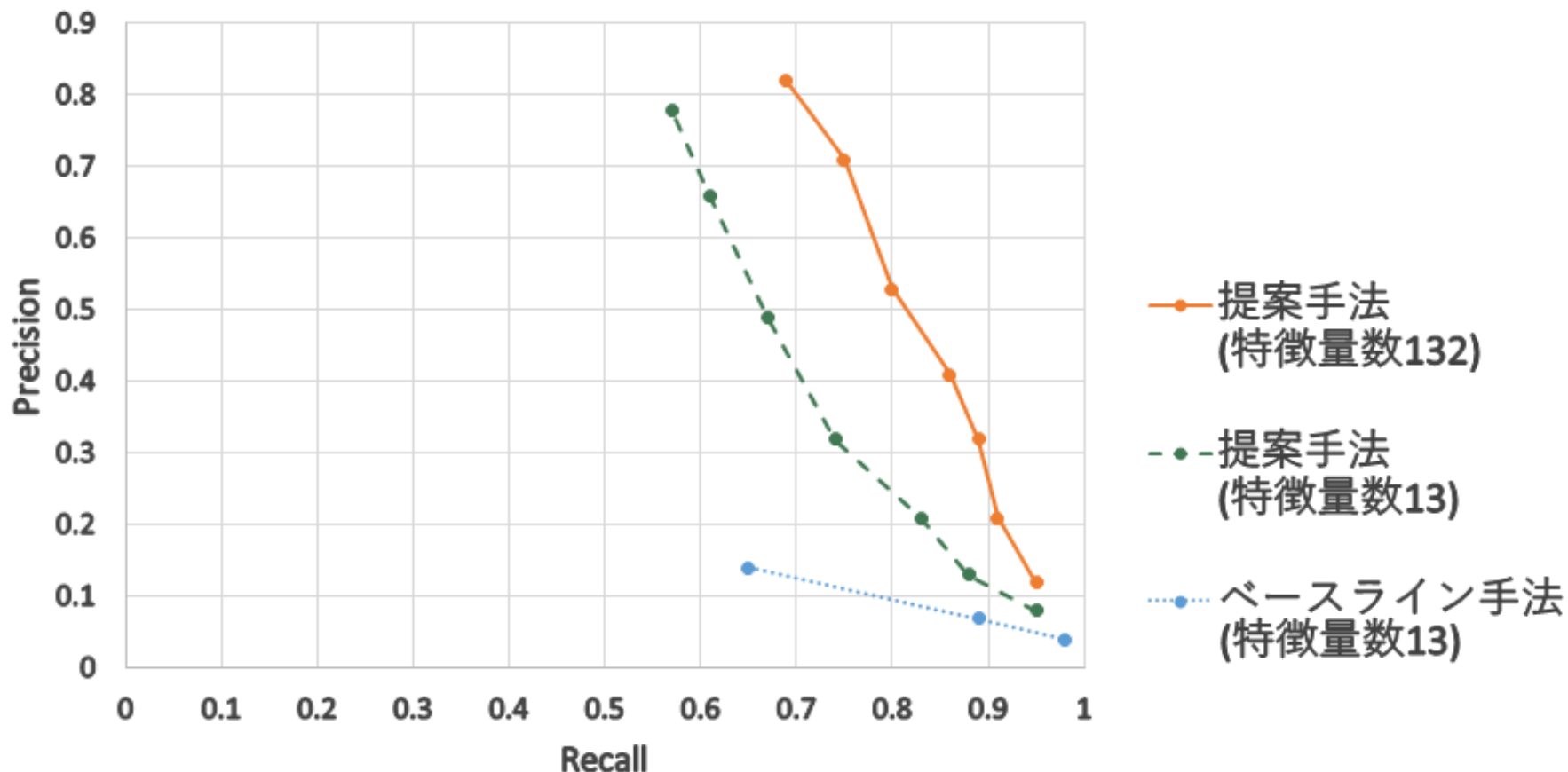


図 提案手法とベースライン手法のPrecision-Recall曲線

- アンダーサンプリングの割合を,
1:16, 1:8, 1:4, 1:2, 1:1, 1:0.5, 1:0.25と変化させてプロット

特徴量の影響度の可視化

- RandomForestClassifierの
feature_importancesというメソッドを使用


Feature ranking:

1. HbA1c (0.220429)
2. 糖代謝判定 (0.169001)
3. インスリン注射または血糖を下げる薬を服用しているか (0.109904)
4. HbA1cの二つ前との差分 (0.091052)
5. 血圧を下げる薬を飲んでいるか (0.056043)
6. HbA1cの二つ前との差分 (0.055971)
7. 尿糖判定 (0.043033)
8. 代表判定 (0.027582)

特徴量の影響度の可視化

- RandomForestClassifierの
feature_importancesというメソッドを使用

Feature ranking:

- 
- 1. HbA1c (0.220429)**
 - 2. 糖代謝判定 (0.169001)**
 - 3. インスリン注射または血糖を下げる薬を服用しているか (0.109904)**
 4. HbA1cの二つ前との差分 (0.091052)
 5. 血圧を下げる薬を飲んでいるか (0.056043)
 6. HbA1cの二つ前との差分 (0.055971)
 7. 尿糖判定 (0.043033)
 8. 代表判定 (0.027582)

影響度ランキングの解釈

➤ HbA1cとは？

ーヘモグロビン中に含まれるグリコヘモグロビンの割合を％で表した値

⇒ **糖尿病の判定**に用いられている

➤ 糖代謝とは？

ー摂取した糖質をエネルギーとして利用したり，脂肪やグリコーゲンとして貯蔵される仕組み

⇒ 糖代謝異常が**糖尿病**につながる

➤ 糖尿病の特徴：血糖値が高い

仮説

- 糖尿病のデータ数 = 921
→ 正例データの73%

複数の対象病名のうち、
糖尿病だけ識別しやすい
のではないか

考察 (糖尿病だけを識別)

- 正例：糖尿病のひとの健診データ
負例：糖尿病以外の重症化病名対象者 + 健康な人
- データサイズ

正例データ	921
負例データ	37998

- Confusion Matrix

		予測されたクラス	
		正例	負例
実際のクラス	正例	836	85
	負例	1812	36186

* 正例のrecall : 0.91
正例のprecision : 0.32

考察 (狭心症だけを識別)

- 正例：狭心症のひとの健診データ
負例：狭心症以外の重症化病名対象者 + 健康な人
- データサイズ

正例データ	229
負例データ	38690

- Confusion Matrix

		予測されたクラス	
		正例	負例
実際のクラス	正例	204	25
	負例	5133	33557

* 正例のrecall : 0.89
正例のprecision : 0.04

まとめ

- 問題設定：「1年以内に重症化するか否か」
- 対象病名全てを正例として識別すると，
recall=0.89, precision=0.32という結果が得られた
- 対象病名の中で，糖尿病は高い精度で識別できる
- 糖尿病以外である狭心症を正例として識別すると，
recall=0.90, precision=0.03にとどまった

今後の課題

- 医師の自由記述欄
⇒ 自然言語処理を施し，特徴量に追加
- 問診票の変化量も特徴量に加える
- 異常検知手法の利用 （不均衡データに適する）
- 糖尿病の指標であるHbA1cを予測のターゲットとする