

## 別添資料3 機械学習を用いた原死因変更有無判定

### はじめに

Irisによる仮原死因が付与できた320,008件中、付帯情報があるデータを対象に、今後の300～500万件規模の研究のための実験を行った。実験では、教師あり機械学習手法であるXGBoostを用いて、Irisが付与した仮原死因が変更されるか否かの二値分類を行った。

### 実験1

#### 方法

##### 手順1. 機械学習用データの生成

「3-2 別添資料2 Figure.1」で示す8\_Shuusei.plを適用して生成されるデータと320,008件で使用された仮原死因のICD-10コードのデータ(1553種類)を組み合わせて、機械学習用のデータを生成する。その仕様は以下の表の通りである。

表1 機械学習用データの仕様

列番号	列名	内容
1	certificateKey	死亡票の識別番号
2	age	年齢(小数点第1位まで)
3	sex	性別(男:1、女:2)
4～1556	各ICD-10	存在しなければ0。存在する場合、 仮原死因に該当:1 I II欄記載有:優先順位に基づき、0.85,0.7,0.55,0.4,0.25
1557～1578	各付帯情報	別添資料3-1<参考資料>の付帯ありに該当すれば1、しなければ0
1579	変更有無	仮原死因が確定原死因と一致すれば1、しなければ0

1579列目に関して、仮原死因はIrisが算出したものであり、確定原死因は人口動態死因オートコーディングシステム(以下、オートコーディングシステム)である。そのため、変更有無は、Irisが算出した仮原死因はオートコーディングシステムが算出する仮原死因と同様であるという前提のもと算出される。したがって、以降の学習結果においても何らかの影響を与える可能性があることに留意しなくてはならない。

##### 手順2. 機械学習の実行

手順1.で生成したデータを用いて機械学習を行う。機械学習では、表1に示す2～1578列目を用いて1579列目の変更有無を予測する。

機械学習の手法として、勾配ブースティング決定木(Gradient Boosting

Decision Tree: GBDT)の一種である XGBoost を使用する。XGBoost は、アンサンブル学習 (複数の弱学習器を組み合わせる手法) のブースティング (弱学習器を逐次的に構築) と決定木 (木構造を用いて分類や回帰を行う手法) を組み合わせたものであり、比較的精度の高い結果を算出することができる。ただし、テストデータの大きさやパラメータの設定が精度に影響を与える可能性が高く、調査が必要である。そこで、実験は以下のように進める。

① テストデータの大きさを以下のように変更し実行

(ア)0.1 (全体の 10%)

(イ)0.15 (全体の 15%)

(ウ)0.2 (全体の 20%)

(エ)0.25 (全体の 25%)

(オ)0.3 (全体の 30%)

ここで、テストデータに用いられなかったデータは学習用の (訓練) データとして用いられ、

② 交差検証とグリッドサーチを用いて実行

交差検証では、データをいくつかのセットに分割したのち、1つをテストデータのセット、残りを学習データとして精度の評価を行う。そして、分割したデータのセットすべてが1回ずつテストデータになるように学習を行なって出てきた精度の平均を算出する。今回は、分割数を5とする。

グリッドサーチとは、学習モデルに用いられるハイパーパラメータ (調整の必要があるパラメータ) を調整する手法であり、指定したハイパーパラメータの全組合せに対して学習を行い、最適なパラメータを選択する。ただし、組み合わせ数が多いほど時間を要するため、今回は決定木の深さにのみ着目し、深さを4,6,8と変えていく。

手順3. 予測と正解の比較

手順 2.から予測した結果をもとに、精度の算出と予測に効果的だとされたパラメータを算出する。

## 結果

① テストデータの大きさ別の結果

各テストデータの大きさ別の結果を図1~図5に示す。各図に関して、左の表は四分表にまとめた結果であり、一番下の行には精度が示されている。また、緑のセ

ルはテストデータの総数、青いセルは各行あるいは各列の総数である。一方、右の表は「変更なし」あるいは「変更あり」に着目した際の感度 (recall)・適合度 (precision)・F 値・特異度 (specificity) を示している。

		正解					
		なし	あり		なし	あり	
予測	変更なし	6792	618	7410	recall	97.3%	47.8%
	変更あり	192	567	759	precision	91.7%	74.7%
		6984	1185	8169	F	0.94	0.58
accuracy:		90.1%			specificity	47.8%	97.3%

図 1 (ア) 0.1 (全体の 10%)

		正解					
		なし	あり		なし	あり	
予測	変更なし	10193	890	11083	recall	97.3%	49.9%
	変更あり	283	888	1171	precision	92.0%	75.8%
		10476	1778	12254	F	0.95	0.60
accuracy:		90.4%			specificity	49.9%	97.3%

図 2 (イ) 0.15 (全体の 15%)

		正解					
		なし	あり		なし	あり	
予測	変更なし	13604	1220	14824	recall	97.4%	48.5%
	変更あり	364	1150	1514	precision	91.8%	76.0%
		13968	2370	16338	F	0.94	0.59
accuracy:		90.3%			specificity	48.5%	97.4%

図 3 (ウ) 0.2 (全体の 20%)

		正解					
		なし	あり		なし	あり	
予測	変更なし	16961	1520	18481	recall	97.1%	48.7%
	変更あり	499	1442	1941	precision	91.8%	74.3%
		17460	2962	20422	F	0.94	0.59
accuracy:		90.1%			specificity	48.7%	97.1%

図 4 (エ) 0.25 (全体の 25%)

		正解					
		なし	あり		なし	あり	
予測	変更なし	20371	1837	22208	recall	97.2%	48.3%
	変更あり	581	1718	2299	precision	91.7%	74.7%
		20952	3555	24507	F	0.94	0.59
accuracy:		90.1%			specificity	48.3%	97.2%

図 5 (オ) 0.3 (全体の 30%)

以上より、得られた精度をテストデータの大きさ別に示すと図 6 のようになる。

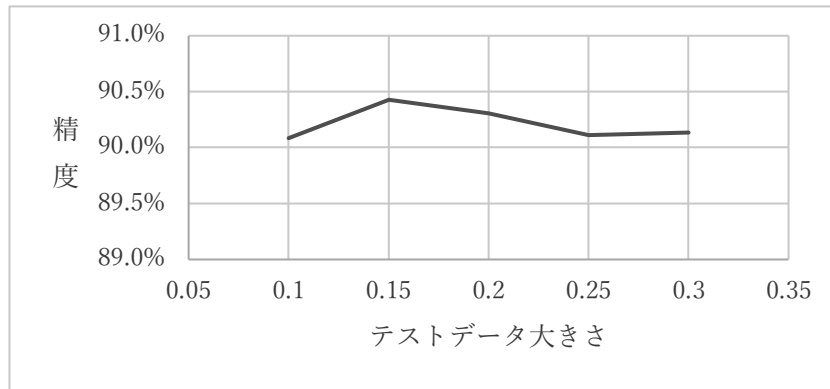


図 6 テストデータの大きさ別の精度推移

② 5 分割交差検証とグリッドサーチを用いた結果

5 分割交差検証とグリッドサーチをした結果、最適な木の深さは 4 であった。結果を図 7 に示す。

		正解		
		なし	あり	
予測	なし	13545	1161	14706
	あり	423	1209	1632
		13968	2370	16338
accuracy:		90.3%		

	なし	あり
recall	97.0%	51.0%
precision	92.1%	74.1%
F	0.94	0.60
specitivity	51.0%	97.0%

図 7 5 分割交差検証 (木の深さ 4) の結果

一方、テストデータの大きさが 0.2 のとき、学習に有効だと考えられた項目を図 8 に示す。

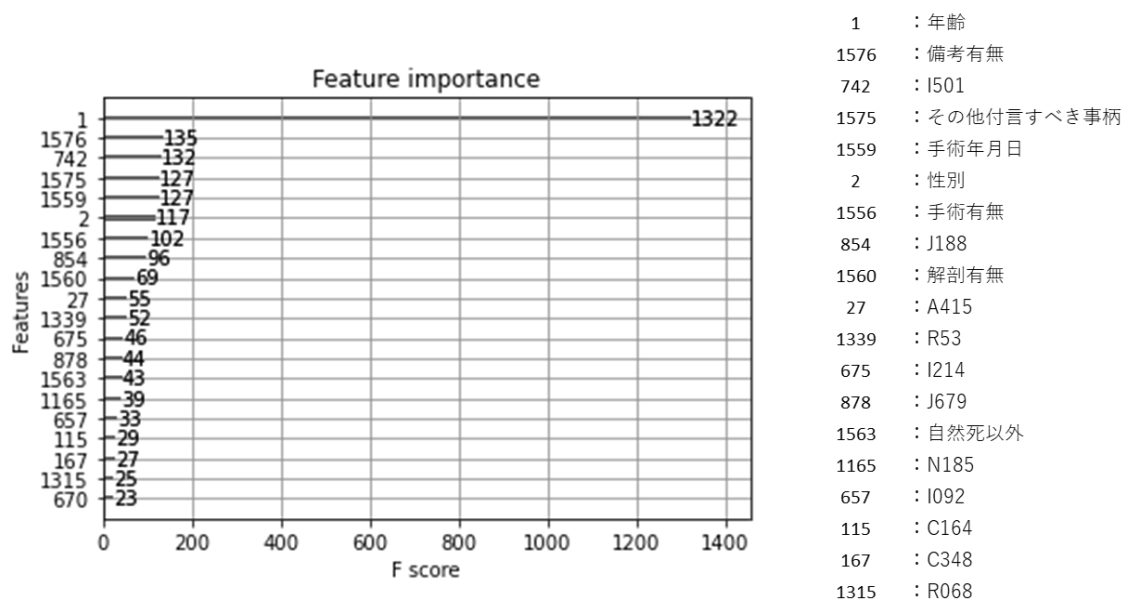


図 8 学習に影響を与えた項目 (テストデータの大きさが 0.2)

## 考察

仮原死因の変更有無に関する機械学習を行った結果、約 9 割の精度で分類が可能であることがわかった。また、図 1~6 に示すテストデータの大きさ別の結果に関して、大きさを 0.1~0.3 まで変化させて実験を行ったが、精度に大きな違いはないと考えられた。また、図 7 に示す 5 分割交差検証とグリッドサーチを行った結果も精度に大きな違いはなかった。そのため、今回の実験ではテストデータの大きさやハイパーパラメータの選択の違いによる学習への影響は少ないと考えられる。

図 8 学習に影響を与えた年齢・性別や ICD-10 コード、付帯情報有無の項目に関して、最も影響を与える項目は年齢であることがわかった。また、そのほかにも備考欄の有無に関してやその他付言すべき事柄に記入があるかなど付帯情報が影響を与えていた。

5 分割交差検証とグリッドサーチにより最適とされたモデルによる予測と正解が異なるものについての考察を行う。

前提として、手順 1.でも述べたように、正解データは Iris (オートコーディングシステムと同じ挙動であると仮定) の算出した仮原死因とオートコーディングシステムによる確定原死因の違いをもとに算出しているため、Iris とオートコーディングシステムの仕様の違いが影響を与えている可能性がある。

また、機械学習による予測で「変更なし」であったが正解は「変更あり」であった場合のうち 108 件 (129 件中) に関しては、Iris が仮原死因を振っていなかった。これは、ICD-10 コードを付与したのちに Iris で仮原死因付与を行う際に、I II 欄に記載された傷病名に対して付与した ICD-10 コードが日本独自コードである場合や、入力された ICD-10 コードに不備 (前期破水は男性には使えないなど、性別や年齢と ICD-10 コードに不一致) がある場合、Iris は仮原死因を付与することができないためである。この問題は、ICD-10 コード付与時に改めて日本独自コードを修正することや Iris が出力するエラー (rejected) のメッセージを参照し仮原死因の付与を行い対策する必要がある。

機械学習による予測で「変更なし」であったが正解は「変更あり」であった場合に関して、各付帯情報の記載がある (1) 件数を算出すると、手術欄の詳細記述欄に記載がある場合が最も多い (686/1161 件) ことがわかった。より精度を上げるためには、詳細記述欄の記述内容に着目する必要がある (特定の単語や文言が仮原死因変更に影響を与える可能性があるため)、次年度以降の課題とする。

一方、機械学習による予測で「変更あり」であったが正解は「変更なし」であった場合に関しても手術欄の詳細記述欄そして手術の有りの番号にマークがされている場合が最も多かった (321/423 件)。機械学習による予測で「変更なし」であったが正解は「変更あり」であった場合と同様に付帯情報の詳細記述欄などを参照する必要がある。ただし、実用する

際は必ず人の目視確認が入るため、数は減らすべきであるが、他の対応を優先するべきである。

#### まとめ

320,008 件のデータを対象に XGBoost を用いた機械学習を行った。機械学習では、記載されている ICD-10 コードと付帯情報の有無に関する項目から、仮死因の変更が行われるか行われなさを予測しており、精度は約 9 割であった。また、その中でも年齢が最も影響を与えていることがわかり、その他の付帯情報の記載有無も有効であることが示唆された。

今回は、データを訓練用とテスト用にのみ分割していたが、今後は、訓練・検証・テストデータに分けて本実験を行う。また、現在は学習で用いる項目が単純な情報の有無にとどまっているが、特に付帯情報の詳細記述に関しては文章を tf-idf などの手法を用いてベクトル化し学習させるなど、より精度を上げるような機械学習用データの生成に取り組む。