匿名化された健康診断と診療履歴の時系列データによる糖尿病罹患予測

清水 正浩 †　石山晴斗 †　菊池浩明 †

明治大学総合数理学部先端メディアサイエンス学科 †

1　はじめに

近年, 機械学習や AI の発展によりビッグデータの利活用が様々な場面で盛んになっている. 個人情報を含んだデータのプライバシー保護のための匿名加工技術が利用・研究されている. 池上ら [1] は 20 万人分の健康診断データと 28 万人分のレセプトデータ, 32 万人分の適用データを使用し, 従来のコホート研究と比較することによって匿名加工情報の有用性を示した. また, 伊藤ら [2] はレセプトデータや健康診断データから得られる個人の身体的特徴や問診表への回答がどの程度一意であるのか調査し, データから個人が識別されるリスクを評価した. しかし, この研究では, 各個人の診療履歴の時系列変化を考慮していない. 身体的特徴量の時系列変化は個人を識別する際に, 有効な情報であると考える.

そこで，本研究では，診療履歴を考慮したデータの未考慮のデータに対する予測精度を比較する. また, 時系列を考慮した際の一意率への影響を考察する.

2　健康診断データの調査と分析方法

2.1　データ

本研究では, DeSC ヘルスケア Inc が取得して匿名加工した健康診断データ, 基本データ, 傷病レセプトデータを使用する.

使用する健康診断データは, 2017 年の健康診断データが存在し, かつ, 2017 年以前に診療履歴がある被験者についての, 体重や身長との身体的特徴 13 属性と問診結果 17 属性の計 30 属性の健康診断結果から成る.

傷病レセプトデータは, 各患者が診断された傷病の記録である. 基本データは被験者の生年月日や性別を示す. そのデータに対して欠損値処理などの前処理を行った.

A prediction model of diabetes based on the anonymized time-series health examination and medical history data

†Masahiro Shimizu, Haruto Ishiyama, Hiroaki Kikuchi, School of Interdisciplinary Mathematical Science, Meiji University.

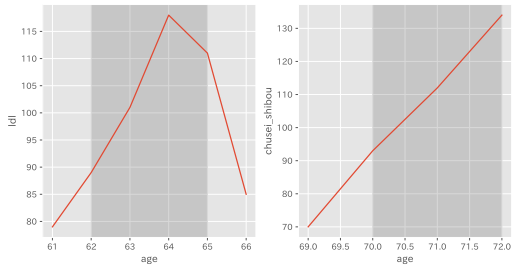

図 1　特徴量の推移の例

2.2　個人の時系列の調査

ランダムサンプリングした 10 名の各特徴量の推移を調査した．図 1 に 2 名の例を示す．糖尿病に罹患した年をグレーで示す．その期間は継続して罹患している．図 1 左は，LDL コレステロール，右は中性脂肪の変化をそれぞれ示す．10 名中 7 名のデータから, 中性脂肪の値の増加が確認された．これらの他にも，血清クレアチニンの値の減少が 10 名中 6 名，腹囲実測値の (+10cm) 増加が 10 名中 7 名にみられた.

2.3　時系列情報付加

健康診断データに対して, 次のように時系列情報を付加する. 特徴量 y の線形単回帰モデルを

$$y = ax + c$$

とする. ここで，x は説明変数（年），a は回帰係数，c は定数項である. 時系列順に図 2 の様に, 線形回帰をして各身体的特徴量の回帰係数 a を基準年のデータセットに付加する. 本アルゴリズムを linear アルゴリズムと呼ぶ.

2.4　機械学習モデルを用いた分析方法

それぞれの健康診断データに対して, 3 年以内に糖尿病罹患を予測するモデルを作成し, 評価を行う. 学習時には罹患者レコードと同数の非罹患者レコードを用いる. サンプリングの際, 非罹患者の分布の違いを考慮するために, 1000 回サンプリングと学習を繰り返した. 学習アルゴリズムにはサポートベクターマシン (SVM) を使用する. 各モデルの評価は, 再現率と適合率の調和平均であ

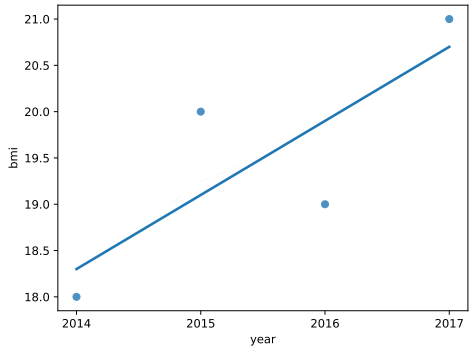

図 2　linear アルゴリズム概要図

る F 値を利用する. モデルは python の scikit-learn を用いて実装する.

表 1　罹患予測するのに用いるデータセット

	被験者数	量的変数属性数	質的変数属性数
3 年以内に糖尿病に罹患する	209	13	17
3 年以内に糖尿病に罹患しない	13673	13	17

2.5　診療履歴の一意率への影響

診療履歴の数に対する安全性への影響を [2] における一意率を用いて観察する. 一意率は, すべての個人の集合における，基準年から length 年間における一意な特徴量ベクトルを持つ個人の割合で定める.

3　分析結果

3.1　SVM についての分析結果

F 値の分布を図**??**, 統計量を表 2 に示す.　linear に関

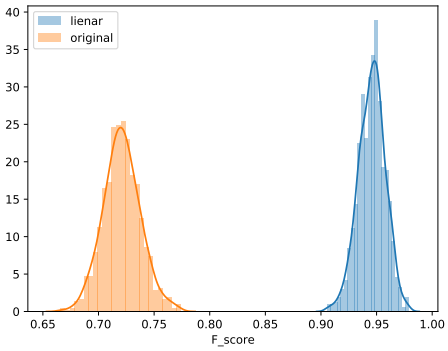

図 3　SVM を用いたときの f 値の分布

しての F 値の平均は 0.9456, 標準偏差は 0.01204 となった. 平均は linear の方が 0.2246 大きくなった. また, 標準偏差は linear の方が小さくなった.

表 2　f 値の統計量

	平均	中央値	標準偏差	最大	最小
original	0.7210	0.7207	0.01678	0.7743	0.6658
linear	0.9456	0.9461	0.01204	0.9790	0.9055

3.2　一意率についての分析結果

表 3 に履歴期間 length に対応した一意率の平均を示す. 各属性は診療履歴の個数を表している. 診療履歴が長

表 3　履歴期間 length における一意率の推移

length	2	3	4	5	6	7
unique rate	0.262	0.424	0.437	0.687	0.783	0.795

くなると, 一意率が単調に増加していくことがわかった.

3.3　考察

結果から, 時系列を考慮した linear の方が高い精度が出ることがわかった. 理由としては, 3 年以内に糖尿病に罹患する者は身体的特徴量の変化の仕方にモデルが識別しやすい特徴量が含まれていると考える. 特に 2.2 節で述べたように, 肥満に関係のある身体的特徴量の変化がモデルの予測に大きく影響を及ぼしていると考える. また, 診療履歴を考慮した一意率は診療履歴が増加していくにつれて共に増加していくということがわかった. ある時点だけに注目すれば一意に識別できないデータでも診療履歴の数が大きくなることによって, これまで同じデータと識別されていたデータが異なるデータと識別されてしまうからだと考えられる.

4　おわりに

本研究では, 時系列データとオリジナルデータの比較を行った. 分析結果より, linear アルゴリズムを用いたデータの F 値が大きいことを確認できた. また, 一意率は診療履歴の数が増加するにつれて共に増加した. 今後の課題として, モデルの精度向上と一意率向上との間にある相関を調査することと考える.

参考文献

[1] 池上，伊藤，菊池, 匿名加工情報の応用 (2): 各種傷病を予測する健康診断モデル，CSS 2020, pp.1230-1237, 2020.

[2] 伊藤，池上，菊池, 匿名加工情報の応用 (1): 健康診断データとレセプトデータの分析とプライバシーリスク評価，CSS 2020, pp.1222-1229, 2020.