別添資料６

厚生労働科学研究費補助金(政策科学総合研究事業（統計情報総合研究事業））
「死因統計の精度及び効率性の向上に資する機械学習の検討に関する研究」
分 担 研 究 報 告 書（令和３年度）

死亡に関わる調査票情報提供に基づいた ICD10 コード自動付与ツールの作成

研究分担者　香川璃奈　（筑波大学医学医療系・講師）

研究要旨

我が国において人口動態調査は国勢調査と並ぶ国の基幹統計であり、中でも死因統計は最も重要な情報の一つである。診療報酬請求や現在普及が進む電子カルテでは標準病名の採用が進められているが、人口動態調査の死因は自由入力病名が元となっており完全な自動集計は困難である。

本研究では、死因確定作業において目視確認に回る理由の中でも、死因を ICD-10 コードに変換できないという点に焦点をあてている。昨年度までに、標準病名と完全一致しない死因の記載を標準病名または ICD-10 コードに変換するルールを作成した上で、そのルールを利用することで、80.06%の症例に ICD10 コードを自動付与できたことを確認した。

本年度はこの結果を利用して、機械学習手法の開発を行った。機械学習を利用する際には、死因の自然文記載をベクトル化、すなわち数学的に利用可能な分散表現に変換する必要がある。ベクトル化の手法として doc2vec を利用したところ、付帯情報の記載によって確定原死因 ICD-10 コードが IRIS による仮原死因コードから変化する症例を、約 94%の正確性で同定できた。

A．研究目的

我が国において人口動態調査は国勢調査と並ぶ国の基幹統計であり、中でも死因統計は最も重要な情報の一つである。診療報酬請求や現在普及が進む電子カルテでは標準病名の採用が進められているが、人口動態調査の死因は自由入力病名が元となっており完全な自動集計は困難である。

我々は令和元年度に平成 27 年～平成 30 年の死亡票とオンライン申請された死亡個票の調査票情報の結合を行なった。結合した情報のことを、以下、突合死亡票 DB(データ数：5,169,031 件)と呼ぶ。これを利用して、標準病名マスターを用いて、全ての Ⅰ欄・II 欄病名に対しほぼ原記載のまま、また助詞、接続詞の除去/展開と言い換えなどの比較的簡便な文字列処理を施すことで、約 65%のⅠ欄・II 欄病名の自動 ICD10 コーディングが可能であるという感触を得た。さらに、Ⅰ欄・II 欄病名を ICD10 コードに変換できたものは約 9 割であった。さらに令和 2 年度は実際の死亡票に記載された病名を ICD10 コードに変換するツールを作成した。独自の対応ルールを利用することで、Ⅰ欄・II 欄病名のすべての自然言語記載病名に ICD10 コードに変換できた件数が全体の 8 割を超えた。

そこで令和 3 年度は、付帯情報に記載されている自然言語記載の内容を数学的に扱うために文(文章)を分散表現で表す手法を利用して、機械学習による死因の ICD10 コード自動付与の精度が変化するか確認した。

さらに、付帯情報が記載されていないにも関わらず iris が付与した原死因と確定原死因が異なる症例が生じる原因を考察することが、将来的な自動での死因 ICD10 コード自動付与ツールの開発に有用な可能性がある。そこで、付帯情報が記載されていないにも関わらず iris が付与した原死因と確定原死因が異なる症例について検討を行なった。

B．研究方法

【実験 1】

自然言語記載の内容を数学的に扱うために文(文章)を分散表現で表す手法を、今後「ベクトル化」と呼ぶ。ベクトル化手法による機械学習精度の違いを検討した。

ベクトル化手法として doc2vec(pvdm)(以下、単に doc2vec と呼ぶ)を利用にした際に、付帯情報の有無によって確定原死因が原死因から変更になるかどうかの 2 値分類を実施し、精度を検討した。

doc2vec の実施プログラムは別添 1 の通りである。

＜＜doc2vec とは＞＞

Doc2vec は任意の長さの文章を固定長のベクトルに変換する技術である。文中の語順や前後に出現する単語を加味したベクトル化が可能である。すなわち、単語をベクトル化した結果の組み合わせだけでは表現できない文章の特徴を表現できる。

Doc2vec にはベクトルに変換するために必要な学習における手法の違いにより 2 種類の手法が存在する[1]。本研究ではそれぞれ利用して結果を比較した。

(1) doc2vec(pv-dm)

文書の id と複数の単語(すなわち、文脈)から直後に続き単語を予測することで文書ベクトルを学習する。

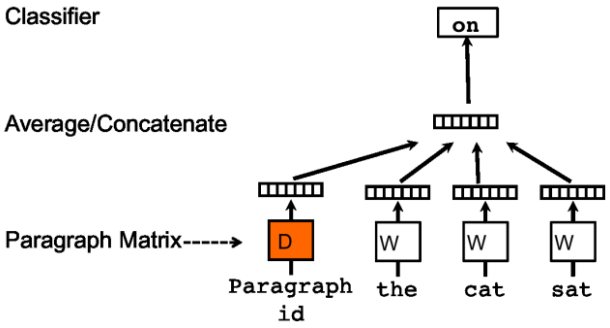

図 1：doc2vec(pv-dm) における学習のイメージ図。 [1]より転載。

(2) doc2vec(pv-dbow)

文書 id のみを入力として、語順を無視して文書に含まれる単語を予測するための学習を行う。単語ベクトル列を学習しないため学習速度が速いとされている。

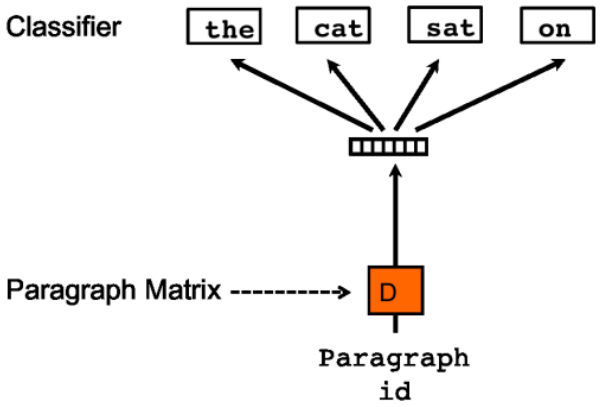

図 2：doc2vec(pv-dbow) における学習のイメージ図。 [1]より転載。

＜＜doc2vec で得られた分散表現を用いた 2 値分類器の学習＞＞

上記 doc2vec で得られた付帯情報に対する分散表現と、共通ベクトル（「性別・年齢、病名の ICD コード、付帯情報の項目の有無」）を結合し、xgboost にて「Iris が決定した仮原死因が変更されるか否か」を予測する 2 値分類器の学習を行った。この詳細については本年度「統括研究報告書」を参照されたい。

【実験 2】

iris が付与した原死因と確定原死因が異なる症例が全部で 17,337 例(iris が付与した原死因と確定原死因の ICD1-コードの組み合わせのユニーク数としては 3,134 種類)存在した。iris が付与した原死因と確定原死因の組み合わせの出現回数の上位 29 件(該当症例 100 例以上の組み合わせ)を目視で確認した。

【実験 1 および 2 に共通すること】

倫理面への配慮

本研究では統計法 33 条に基づき申請したデータを利用した。申請の通り、インターネットに繋がない端末上でのみデータの閲覧作業を行うことで個人情報に配慮した。

C．研究結果

【実験 1】

精度は以下の通りであった。

表 1：ベクトル化手法の違いに基づく機械学習精度の違い。

評価尺度	Embedding 手法	2020	2019	2018	2017
Accuracy	DOC2VEC (PV-DM)	0.933	0.930	0.931	0.930
	DOC2VEC (PV-DBOW)	0.944	0.941	0.942	0.943
ROC-AUC	DOC2VEC (PV-DM)	0.933	0.929	0.932	0.930
	DOC2VEC (PV-DBOW)	0.944	0.943	0.943	0.946
PR-AUC	DOC2VEC (PV-DM)	0.782	0.771	0.785	0.780
	DOC2VEC (PV-DBOW)	0.830	0.821	0.831	0.837

この結果から、doc2vec(pv-dm)と doc2vec(pv-dbow)はいずれも高い精度を示すこと、doc2vec(pv-dbow)の方が僅かに高い精度を示すことを確認できた。

【実験 2】

iris が付与した原死因と確定原死因の ICD10 コードの間の関係性を、以下の 6 通りに分類した。

(A)医学的にはほぼ同義

(B)確定原死因が iris 原死因の下位概念（病変部位が特定されている、など）

(C)iris 原死因が確定原死因の下位概念（病変部位が特定されている、など）

(D)確定原死因が原因となって iris 原死因が生じたと想定される症例

(E)iris 原死因が原因となって確定原死因が生じたと想定される症例

(F)その他

なお以下の表 1 における iris 原死因の自然言語記載病名は各 ICD10 コードについて死亡表において出現回数が最も多い記載を中心として選択したものである。

この結果から、iris 原死因と確定原死因が異なる場合に、iris 原死因と全く異なる確定原死因が付与される事例はほぼ存在しないことを確認できた。

また表 1 に記載されている事例について、コーディングマニュアルに記載されている、原死

因コーディングのための注や連鎖表も確認したが、考察に活用できる知見を見つけることはできなかった。

表 2：iris が付与した原死因と確定原死因の ICD10 コードが異なる組み合わせ

確定原死因	Iris 原死因	該当個数	確定原死因名	Iris 原死因名	関係性の分類
K566	K567	707	Ｓ状結腸狭窄症	亜イレウス	B
I639	I638	504	虚血性脳卒中	出血性脳梗塞	F
C240	C248	460	下部胆管癌	#N/A	F
J189	F03	382	急性肺炎	原発性認知症	E
G309	G301	376	アルツハイマー型認知症	アルツハイマー型老年認知症	A
J189	J440	328	急性肺炎	下気道感染を伴う慢性閉塞性肺疾患	C
I252	I258	299	陳旧性下壁心筋梗塞	冠状動脈炎	E
I693	I639	264	小脳梗塞後遺症	虚血性脳卒中	B
J189	G20	251	急性肺炎	一側性パーキンソン症候群	E
J449	J440	245	慢性閉塞性肺疾患	下気道感染を伴う慢性閉塞性肺疾患	C
R99	空白	237	原因不明の死亡	#N/A	F
J189	G301	221	急性肺炎	アルツハイマー型老年認知症	E
E149	E146	202	糖尿病・糖尿病性合併症なし	高血糖高浸透圧症候群	D
I638	I635	193	出血性脳梗塞	延髄梗塞	C
J849	J841	168	間質性肺炎	炎症後肺線維症	D
I639	I635	163	虚血性脳卒中	延髄梗塞	C
J189	J439	153	急性肺炎	萎縮性肺気腫	D
J690	F03	152	胃分泌物嚥下性肺炎	原発性認知症	E
E142	E147	150	キンメルスチール・ウイルソン症候群	#N/A	F
I219	I258	147	ＳＴ上昇型急性心筋梗塞	冠状動脈炎	E
I500	J189	146	右室不全	急性肺炎	F
A319	B948	135	非結核性抗酸菌症	ジフテリア後麻痺	
R54	I509	135	老化	急性心不全	D
G318	G239	129	ＨＤＬＳ	基底核変性症	F
C220	K746	122	肝癌	萎縮性肝硬変	E
C928	C920	121	骨髄異形成関連変化を伴う急性骨髄性白血病	ＲＡＥＢ－ｔ	C
C220	B182	111	肝癌	Ｃ型肝炎	E
N189	I509	111	慢性腎臓病	急性心不全	D or E
S065	I620	100	外傷性硬膜下水腫	若年性慢性硬膜下血腫	C

D．考察

実験 1 の結果から、付帯情報により、iris が付与した原死因名と確定原死因が異なるかを機械学習で分類する問題について、doc2vec(pv-dbow)が有効であることを確認した。

また、実験 2 については、将来的に、このようなインストラクションマニュアルに記載されていないコーディングのルールが明文化されることで、自動での死因 ICD10 コード自動付

与ツールの開発にも有用であると考える。

E．結論

本年度研究では、付帯情報により、iris が付与した原死因名と確定原死因が異なるかを機械学習で分類する問題について、doc2vec が有効であること、また pv-dm と pv-dbow の２種類では pv-dbow の方が有効であることを確認した。

F．健康危険情報

なし

G．研究発表

なし

H．知的財産権の出願・登録状況

なし

参考文献

[1]Le, Quoc, and Tomas Mikolov. ″Distributed representations of sentences and documents.″ International conference on machine learning. PMLR, 2014.

本研究で用いたプログラムソースは、本報告書全体の添付資料「**機械学習用データセット作成プログラムソース Doc2Vec (PV-DM / PV-DBOW)**」を参照されたい。