テキストマイニングによる 退院サマリー自動分類の試み

長瀬　永遠

富山県立大学　情報基盤工学講座

December 4, 2020

はじめに

本研究の目的

電子化され, 大量に蓄積された医療文書に対してテキストマイニングを行うことで, 医療文書の 2 次利用可能性を示す.

研究のながれ

1. 医療文書に対してテキストマイニングを行うための索引の作成
2. 各疾患における退院サマリーの特徴抽出
3. 抽出した特徴をもとに, 退院サマリーから疾患名が特定できるかを検証
4. 特定した例に対してデンドログラムを作成し, 分類を視覚化

対象

千葉大学附属病院病院情報システムに保存されている 36,335 症例の退院サマリー

⇒ 症例数が 100 以上ある 50 疾患を算出

⇒ 各臓器の代表的疾患である 13 疾患を選定

Table 1: 各臓器の代表 13 疾患とその症例数

疾患名	臓器	ICD-9	症例数
胃悪性新生物	消化器	151	524症例
肝，肝内胆管の悪性新生物	肝臓・胆	155	483症例
気管・気管支の悪性新生物	呼吸器	162	687症例
乳房の悪性新生物	乳房	174	363症例
前立腺悪性腫瘍	男性器	185	340症例
腎臓の悪性新生物	腎臓	189	158症例
リンパおよび組織球組織の悪性新生物	血液	202	153症例
糖尿病	内分泌	250	293症例
統合失調症	精神	295	104症例
白内障	眼	366	777症例
喘息	アレルギー	493	114症例
瘢痕拘縮	皮膚	709	133症例
変形性関節症	運動器	715	188症例

解析環境の構築

形態素解析

- 使用したシステム：「茶筅」（奈良先端科学技術大）
- 追加した辞書：MEID 辞書（医学辞書）

辞書の再構築

臨床現場にて作成された文書には, 略語が多用される.

⇓

退院サマリーを形態素解析し, MEID 辞書にない単語から各疾患における出現頻度上位 50 位までの用語を辞書に追加.

特徴の抽出方法

退院サマリーベクトル

D：対象とする文書集合, $d_1, d_2, \cdots, d_j, \cdots, d_n$：退院サマリー,
$w_1, w_2, \cdots, w_i, \cdots, w_m$：$D$ から抽出された索引語,
α_{ij}：ある退院サマリー d_j におけるある索引語 w_i の重み

$$\overrightarrow{d_j} = [\alpha_{1j}\ \alpha_{2j} \cdots \alpha_{ij} \cdots \alpha_{mj}]^T \tag{1}$$

$\Rightarrow$ 退院サマリーベクトル

索引語文書行列

$$\overrightarrow{D} = [\overrightarrow{d_1}\ \overrightarrow{d_2} \cdots \overrightarrow{d_n}] = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{pmatrix} \tag{2}$$

$\Rightarrow$ 索引後文書行列

重みの算出方法

tf×idf 法

α_{ij}：重み,

$$\alpha_{ij} = \frac{l_{ij} g_i}{n_j} \tag{3}$$

f_{ij}：w_i の d_j における出現頻度
n：対象とする退院サマリーの症例数
n_i：対象の退院サマリーにおける w_i を含む症例数

$$l_{ij} = \log(1 + f_{ij}) \tag{4}$$

$$g_i = \log(\frac{n}{n_i}) \tag{5}$$

$$n_j = \sqrt{\sum_{i=1}^{m} (l_{ij} g_i)^2} \tag{6}$$

貪欲アルゴリズムによるデンドログラムの作成

最大木問題

各ノード（今回は疾患）をつなぐ木の重み（今回は類似度）の総和が最大となる木を求める問題.

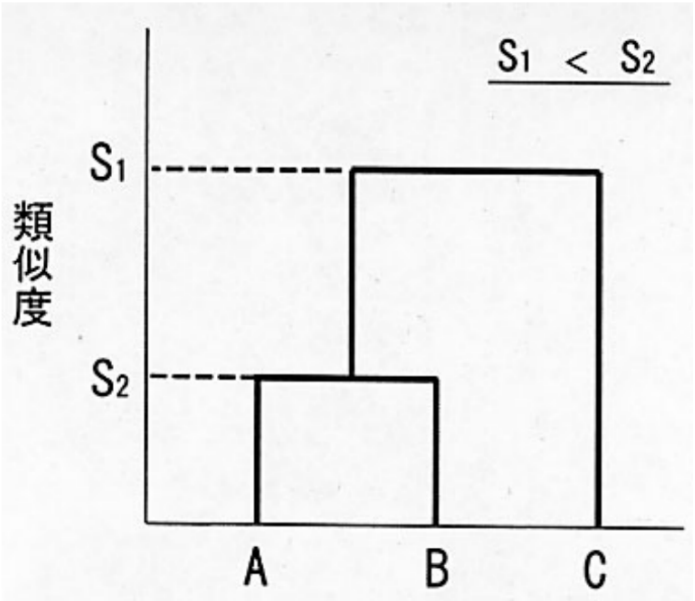

図 1: デンドログラムの例

13 疾患の重要語の抽出

索引語の抽出

茶筅を用いて 13 疾患の退院サマリーの形態素解析
⇓ 結果
7,918 語の索引語が抽出

Table 2: 重み上位 10 位の索引語 1

	胃癌		肝癌		肺癌		乳癌		前立腺癌	
	索引語	α	索引語	α	索引語	α	索引語	α	索引語	α
1	前庭	0.104	エタノール	0.092	右中葉	0.082	乳管	0.195	サドルブロック	0.154
2	胃体	0.101	PHA	0.09	扁平上皮癌	0.078	C領域	0.172	前立腺	0.149
3	フード	0.099	コイル	0.088	気管分岐部	0.075	乳腺症	0.164	前立腺全摘除術	0.146
4	胃透視	0.092	前枝	0.086	肺瘻	0.071	胸筋	0.159	骨盤リンパ節	0.134
5	SE	0.089	Fc	0.083	入口	0.069	ノルバデックス	0.141	PK	0.121
6	胃全摘術	0.089	食道静脈瘤	0.08	肺機能	0.068	上肢挙上	0.138	除睾術	0.117
7	GFS	0.087	門脈	0.079	葉間	0.067	マンモグラフィー	0.125	ホンバン	0.109
8	亜全摘	0.083	アミノレバン	0.077	プラシ	0.067	癌検診	0.121	タンデム	0.104
9	胃切除術	0.081	右枝	0.077	壁側胸膜	0.067	大胸筋	0.112	側精巣	0.104
10	器械	0.079	完全壊死	0.077	腹様部	0.067	乳房	0.111	直腸出血	0.104

	腎癌		悪性リンパ腫		糖尿病		統合失調症		白内障	
	索引語	α	索引語	α	索引語	α	索引語	α	索引語	α
1	腎腫瘍	0.167	PUVA	0.108	補食	0.11	幻聴	0.122	点眼液	0.211
2	腎盂	0.144	可溶性	0.102	腎症	0.091	拒絶	0.107	眼	0.205
3	尿管腫瘍	0.132	髄注	0.1	硝子体出血	0.089	疎通性	0.101	ミドリンP	0.173
4	右腎盂	0.126	幹細胞	0.097	神経伝導速度	0.087	隔離	0.098	水晶体乳化	0.168
5	腎細胞癌	0.12	耳下腺	0.095	ケトン	0.086	被害妄想	0.095	両眼	0.156
6	右尿管口	0.112	悪性リンパ腫	0.092	グルカゴン	0.082	妄想	0.094	乳化	0.148
7	腎摘除術	0.107	リンパ腫	0.087	ケトン体	0.081	行為	0.093	眼内レンズ	0.142
8	腎部分切除術	0.104	右扁桃	0.083	強化療法	0.079	空笑	0.089	吸引術	0.137
9	拡大率	0.103	躯幹	0.083	マイクロゾーム	0.077	妄想状態	0.087	右眼	0.136
10	上極	0.103	上咽頭	0.079	肥満度	0.072	ダール	0.085	左眼	0.132

Table 3: 重み上位 10 位の索引語 2

	喘息		瘢痕拘縮		変形性関節症	
	索引語	α	索引語	α	索引語	α
1	インタール	0.192	瘢痕拘縮	0.189	裂隙	0.14
2	陥没	0.12	エキスパンダー	0.179	下腿周径	0.125
3	スギ	0.118	プロテーゼ	0.146	内反	0.12
4	ダニ	0.118	植皮術	0.143	荷重	0.118
5	胸骨上窩	0.112	シリコン	0.142	骨棘形成	0.113
6	呼気性喘鳴	0.11	挫創	0.134	左膝	0.112
7	持続吸入	0.108	ケロイド	0.118	CE角	0.11
8	大発作	0.105	左上眼瞼	0.118	外反	0.11
9	ブタクサ	0.103	全層植皮	0.116	脚長差	0.11
10	湿性	0.101	修正	0.112	動揺性	0.11

退院サマリーからの疾患の特定

特徴抽出に用いたのとは別の 13 疾患の退院サマリーを各 30 症例, 計 390 症例無作為に抽出し, 既出の 13 疾患における退院サマリーベクトルとの内積を計算.
⇒13 疾患に対する類似度を, 症例ごとにレーダーチャートで表現.

Table 4: 各疾患の正診率

疾患名	臓器	ICD-9	診断と一致	複数疾患の疑いあり	診断と異なる判定	判定不明
胃癌	消化器	151	24 / 30	4 / 30	0 / 30	2 / 30
肝癌	肝臓・胆嚢	155	20 / 30	2 / 30	0 / 30	8 / 30
肺癌	呼吸器	162	19 / 30	3 / 30	0 / 30	8 / 30
乳癌	乳房	174	19 / 30	1 / 30	1 / 30	9 / 30
前立腺癌	男性器	185	25 / 30	3 / 30	0 / 30	2 / 30
腎癌	腎臓	189	21 / 30	2 / 30	1 / 30	6 / 30
悪性リンパ腫	血液	202	17 / 30	6 / 30	1 / 30	6 / 30
糖尿病	内分泌	250	19 / 30	7 / 30	2 / 30	2 / 30
統合失調症	精神	295	28 / 30	1 / 30	0 / 30	1 / 30
白内障	眼	366	28 / 30	2 / 30	0 / 30	0 / 30
喘息	アレルギー	493	27 / 30	0 / 30	0 / 30	3 / 30
瘢痕拘縮	皮膚	709	13 / 30	0 / 30	4 / 30	13 / 30
変形性関節症	運動器	715	30 / 30	0 / 30	0 / 30	0 / 30
計			290 / 390 (74%)	31 / 390 (8%)	9 / 390 (2%)	60 / 390 (15%)

13 疾患のデンドログラム

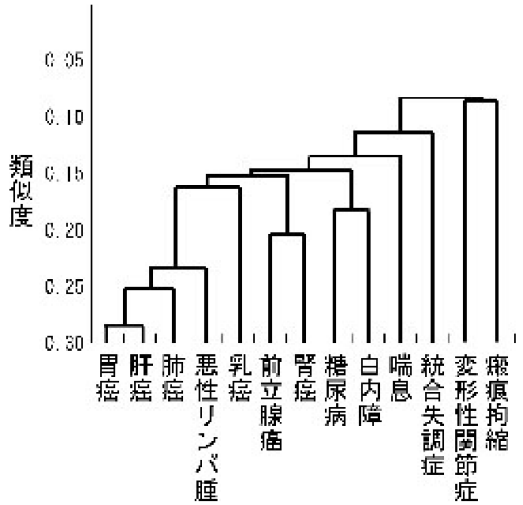

図 2: 13 疾患の類似度の最大全域木によるデンドログラム

- 胃癌と肝癌の類似度が最も高い.
  ⇒ 両疾患とも悪性腫瘍ということから妥当.
- 変形性関節症と瘢痕拘縮のグループの類似度が最も低い.

おわりに

今後の課題

- 解析に用いる辞書の改良
- 扱うデータを増やすことによる，退院サマリーベクトルの改善
- 重み付け方法の再考
- より多数の疾患を適用したデンドログラムの妥当性の確認

今後の展望

- 既存の疾患の退院サマリーベクトルとの類似度をもとにした，新たな疾患の発見（退院サマリーベクトル）
- 合併症の可能性の示唆（レーダーチャート）
- 疾患群同士の相関関係の視覚化による新たな知見の誘発（デンドログラム）