医療テキストデータに対するトピックモデルの応用

大谷　誠[1)]、松田　晋哉[1,2)]

1）産業医科大学　産業保健データサイエンスセンター
2）産業医科大学　医学部　公衆衛生学教室

要旨：本研究では医療データであるDPCデータを用いテキストマイニングを行った。医療データは数値データやカテゴリー尺度、ICD10といったコードなど分析し医療行為にフィードバックが行えるよう精緻化されてきている。DPCコード毎のトピックの出現度をみると、手術・処置有無に関してリハビリは前回手術を行ったDPCコード、症状不良は前回入院時に生検法をした化学療法と放射線治療の両方を行っているDPCコードに関連していた。また、化学療法と放射線治療の両方が行われているDPCコードは3から4つのトピックを含んでいた。本研究によりテキスト分析により医療行為内容の体系的把握が可能になる可能性が示された。医療行為の均一化や医療職の負担軽減を実現するために、より精度の高いテキスト分析の手法を開発していく必要がある。

キーワード：　DPCデータ、テキストマイニング、トピックモデル

1．はじめに

平成30年8月戦略的イノベーション創造プログラム第2期においてAIホスピタルによる高度診断・治療システムが掲げられている。AIホスピタルによるシステムは「高度で先進的な医療サービスを提供するとともに、医療機関における効率化を図り、医師や看護師などの医療従事者の抜本的な負担の軽減を実現する。」ものである。システム構築に向け2020年度までの到達目標の1つに「セキュリティの高い医療情報データベースの構築・医療有用情報抽出技術の開発」が掲げられている[1)]。

医療情報には血液検査やバイタルサインなどの数値データ、CTやMRIなどの画像データ、SOAPなどで電子カルテに記入されるテキストデータなどがある。画像データに関してはG. Litjensの論文によると2016年以降にDeep Learningを用いた画像解析に関する論文等が発表されていると報告されている[2)]。

情報抽出の技術としてデータマイニングがある。データマイニングとは「データの中に埋め込まれている知識の発掘」である[3)]。データマイニングの分析対象がテキストデータとなったものがテキストマイニングである。医療分野でのテキストマイニングは1960年代から行われている[4)]。日本では退院サマリーからの疾患名の特定[5)]や看護記録からの有用情報の抽出[6)]に関する研究が行われている。医療分野でテキストデータが入力されているものとしてDPC（Diagnosis Procedure Combination）データがある。

DPCデータは患者の基本情報や入退院先に関する情報、疾患の重症度など診療録情報が記載される様式1、実施・処方される手術や処置、薬剤の情報など診療報酬請求情報が記載される入院EF統

合ファイルなどで構成されている[7]。様式1には再入院および再転棟に関する自由記載の入力欄がある。本研究では再入院に着目しそのうち計画外の再入院のテキストデータから再入院する患者の傾向をテキストマイニングで抽出することを目的に分析を行う。

2．方法

2.1　分析データおよび分析対象

一般社団法人診断群分類研究支援機構が収集した平成26年度及び平成27年度のDPCデータを用いた。分析対象としてDPC6桁コード（病名）が「肺の悪性腫瘍（040040）」の患者を選択した。DPCデータの再入院には計画的再入院、予期された再入院、予期せぬ再入院の3種類が用意されている。本研究では「予期された再入院および予期せぬ再入院」の患者を抽出し、自由記述欄に記載がある患者（42,605名）を対象とした。

2.2　分析方法

本研究ではテキストマイニングの手法にトピックモデルを用いた。トピックモデルとは文書に出現する単語とその出現回数の情報からそれぞれの文書に潜在的に存在するトピックを推定する手法である[8]。本研究ではトピックモデルの手法の1つである潜在的ディレクレ配分法（Latent Dirichlet Allocation：LDA）[9]を用いた。トピックモデルによるテキストマイニングの概要を図1にLDAのトピック生成過程のアルゴリズムを以下に示す。本研究ではトピック数をK=7とした。

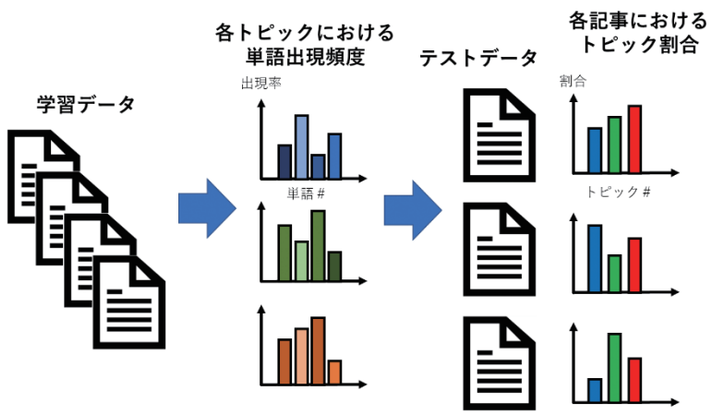

図1　トピックモデルの概要

2.3　学習データ

トピックモデルに学習させるデータの作成は以下の手順で行った（図2参照）。

① 入院患者毎に前回入院時のDPCコード（14桁）と再入院理由の自由記載を抽出

② DPCコード毎に自由記載を結合（前処理として句読点や記号の除去、半角全角や大文字小文字の統一を行った）

③ 結合後の文章に対し形態素解析を行い名詞（一般、固有名詞、形容動詞語幹）と形容詞（自立）の単語を抽出（形態素解析にはPythonのパッケージであるJanomeを用いた）

アルゴリズム 1 生成アルゴリズム

for トピック $k = 1, \cdots, K$ do
　単語分布を生成 $\phi_k \sim \text{Dirichlet}(\beta)$
end for
for 文書 $d = 1, \cdots, D$ do
　トピック分布を生成 $\theta_d \sim \text{Dirichlet}(\alpha)$
　for 単語 $n = 1, \cdots, N_d$ do
　　トピックを生成 $z_{d,n} \sim \text{Categorical}(\theta_d)$
　　単語を生成 $w_{d,n} \sim \text{Categorical}(\theta_{z_{d,n}})$
　end for
end for

①データ抽出

患者ID	DPCコード	自由記載
01	040040xx9903xx	TEXT1
02	040040xx9903xx	TEXT2
03	040040xx9901xx	TEXT3
04	040040xx9904xx	TEXT4
05	040040xx9904xx	TEXT5
06	040040xx9900xx	TEXT6
07	040040xx9901xx	TEXT7
08	040040xx9904xx	TEXT8
09	040040xx9900xx	TEXT9
10	040040xx9901xx	TEXT10
11	040040xx9900xx	TEXT11
12	040040xx9901xx	TEXT12
13	040040xx9901xx	TEXT13
14	040040xx9904xx	TEXT14
15	040040xx9900xx	TEXT15

②DPCコードごとにデータを結合

DPCコード	自由記載結合
040040xx9900xx	TEXT6 TEXT9 TEXT11 TEXT15
040040xx9901xx	TEXT3 TEXT7 TEXT10 TEXT12 TEXT13
040040xx9903xx	TEXT1 TEXT2
040040xx9904xx	TEXT4 TEXT5 TEXT8 TEXT14

③文章を形態素に分割

DPCコード	形態素解析
040040xx9900xx	w23 w2 w13 w48 w46 w6 w5 w24 w25 w40 w46 w22 w47 w13 w45 w41 w42 w34 w50 w10 w2 w35 w5
040040xx9901xx	w24 w38 w35 w11 w48 w44 w11 w39 w22 w27 w4 w39 w29 w41 w14 w4 w13 w20 w14 w50 w22 w50 w24 w5 w11 w7 w23 w21 w50 w32 w28 w12
040040xx9903xx	w37 w43 w1 w31 w15 w4 w5 w49 w49 w47 w14 w1 w48 w9 w25 w40 w12 w43 w27 w24 w46 w28
040040xx9904xx	w34 w12 w48 w26 w23 w45 w7 w43 w25 w29 w46 w5 w18 w4 w14 w8 w17 w48 w4 w1 w41 w10 w39 w34

図 2　学習データ作成フロー図

3．結果

トピックモデルによる各トピックと単語の関係を図 3 および表 1 に示す。図 3 は各トピックで出現する単語の頻度を文字の大小で表現したものである。表 1 は各トピックに対して出現する単語からトピック名をラベル付けしたものである。DPC コード毎のトピックの出現度を表 2 に示す。Topic#2 のリハビリに関しては前回入院時に手術を行ったコードのみ表れた。Topic#6 の症状不良に関しては前回入院時に生検法をされており化学療法と放射線治療の両方を行っているコードのみに表れた。また化学療法と放射線治療の両方が行われているコードは 3 から 4 つのトピックが含まれていた（040040xx97x4xx、040040xx9904xx、040040xx9914xx）。Topic#5 の治療に関しては全コード中の約 8 割、Topic#0 の検査に関しては約 4 割で現れた。

図３　トピック内の単語出現頻度

表１　トピック名

Topic#	トピック名	係数＊単語
0	検査	0.009* 気管支 +0.007* 目的 +0.006* 肺癌 +0.006* 方針 +0.006* 前回 + 0.005* 医療 +0.005* 放射線 +0.005*PCLB+0.004* 必要 +0.004* 検査
1	治療不良	0.007* 副作用 +0.006* 繰り返し +0.006* 小腸 +0.005* 化学 +0.005* 病状 + 0.005* 悪い +0.005* 調子 +0.004* イレッサ +0.004* 疾患 +0.004* 治療不良
2	リハビリ	0.006* アナフィラキシーショック +0.005* 筋力 +0.005* 胃癌 +0.004* 肺癌 + 0.004* 病理 +0.004* リハビリ +0.004* 目的 +0.004* 食欲 +0.004* 方針
3	他疾患	0.007* 肺癌 +0.006* 肺炎 +0.006* 胸膜炎 +0.005* 肺結核 +0.005* 腹痛 + 0.005* 腹膜炎 +0.005* 目的 +0.005* 抗生 +0.005* 転院 +0.004* 他疾患
4	緊急	0.004* 方針 +0.004* 目的 +0.004* 救急 +0.003* 緊急 +0.003* 化学 + 0.003* 頭蓋 +0.003*RFA+0.003* ロキソプロフェン +0.003* 内圧
5	治療	0.010* 化学 +0.008* 療法 +0.008* 良好 +0.007* 外来 +0.007* 目的 + 0.005* アナフィラキシーショック +0.005* 緊急 +0.005* 狭窄 +0.005* 放射線 + 0.005* 治療
6	症状不良	0.009* 悪心 +0.006* 骨髄 +0.003* 不振 +0.003* 食欲 +0.002* 外来 + 0.002* 前回 +0.002* 化学 +0.002* 療法 +0.001* ドレーン +0.001* 症状不良

表 2　DPC コードとトピックの関係

DPC14	Topic #0	Topic #1	Topic #2	Topic #3	Topic #4	Topic #5	Topic #6	手術	処置 1	処置 2
040040xx97x0xx	0.37		0.12	0.02		0.49		○	×	×
040040xx97x10x				0.86		0.14		○	×	○
040040xx97x2xx	0.90			0.10				○	×	○
040040xx97x3xx		0.56				0.44		○	×	○
040040xx97x4xx	0.05	0.05			0.17	0.72		○	×	○
040040xx97x5xx	1.00							○	×	○
040040xx97x6xx			0.57			0.43		○	×	○
040040xx97x7xx						1.00		○	×	○
040040xx97x80x						1.00		○	×	○
040040xx97x81x						1.00		○	×	○
040040xx9900xx	0.18			0.39		0.43		×	×	×
040040xx9901xx				0.99				×	×	○
040040xx9902xx	0.26				0.41	0.34		×	×	○
040040xx9903xx						1.00		×	×	○
040040xx9904xx	0.03	0.01		0.05		0.91		×	×	○
040040xx9905xx		0.73				0.27		×	×	○
040040xx9906xx						1.00		×	×	○
040040xx9907xx		0.29				0.70		×	×	○
040040xx9908xx	0.41					0.59		×	×	○
040040xx99100x	0.89					0.11		×	○	×
040040xx99101x	0.78					0.22		×	○	×
040040xx9911xx	1.00							×	○	○
040040xx9912xx				0.68		0.32		×	○	○
040040xx9913xx						0.68	0.31	×	○	○
040040xx9914xx	0.22	0.43				0.36		×	○	○
040040xx9915xx						1.00		×	○	○
040040xx9916xx				0.51		0.49		×	○	○
040040xx9917xx		1.00						×	○	○

4．考察

テキストマイニングの対象として再入院理由を選択した。再入院率は医療の質の評価指標[10]として用いられており、再入院率を下げることが期待されている。本研究では計画的再入院以外の再入院を分析対象とした。現在、治療と仕事の両立支援[11]が国から各企業に求められており就労者にとっても計画的に治療が進むことで就労への障害が減りより積極的な治療が望めると思われる。本研究では肺

の悪性腫瘍を対象とした。2018 年度より診療報酬としてがん患者対象に療養・就労両立支援指導料が始まったこと、がんの中の罹患者数が 3 番目に多く死亡数が一番多い[12]ことから対象の疾患とした。再入院にあたり前回入院における医療行為により理由のパターンがあると仮説を立て DPC14 桁別に分析を行ったがパターンは見つけることができなかった。その要因として 2 つ考えられ①用語のばらつきと②各施設の入力内容の粒度の違いが考えられる。①に関しては入力される用語は日本語入力や英語入力、略語などバリエーションがあった。例えばリハビリテーションは「リハビリテーション」、「リハビリ」、「リハ」である。この対策としては用語の対応表を作成し正式名称に変換後、分析する必要があると考えられる。②に関しては丁寧に入力している施設もある一方、2 文字しか入力していない施設があった。データを作成したテキストは自由記載欄ではあるが、知識の共有や分析への応用を考えると最低でも「文」として入力を求める必要があると思われる。本研究ではトピックに割り当てられた単語から研究者がトピック名の決定を行った。トピック名決定には専門的な知識が必要でありデータによって異なる専門家が必要になると思われる。

本研究では医療データである DPC データを用いテキストマイニングを行った。医療データは数値データやカテゴリー尺度、ICD10 といったコードなど分析し医療行為にフィードバックが行えるよう精緻化されてきている。しかし電子カルテ内にはテキストデータが多く存在していることも事実でありその中に共有すべきノウハウが存在している。そのためテキストからノウハウを抽出し共有することで医療行為の均一化や医療職の負担軽減に寄与すると考えられることから、より精度の高いテキスト分析の手法を開発していく必要があると思われる。

参考文献

1）内閣府，戦略的イノベーション創造プログラム（SIP）の概要　研究開発計画　第 2 期　12 課題，https://www8.cao.go.jp/cstp/gaiyo/sip/kenkyugaiyo2.pdf，2019 年 12 月 10 日閲覧

2）Geert Litjens, et al., A survey on deep learning in medical image analysis, Medical Image Analysis, Vol. 42, pp. 60-88, 2017

3）人工知能学会編，人工知能学辞典，協立出版株式会社，2005

4）Hercules Dalianis, Clinical Text Mining Secondary Use of Electronic Patient Records, Springer, 2018

5）小野大樹ら，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，医療情報学，24(1)，pp 35-44，2004

6）村松洋，看護記録のテキストマイニング，情報処理学会論文誌データベース，Vol. 3，No. 3，pp. 112-122，2010

7）平成 30 年度「DPC 導入の影響評価に係る調査」実施説明資料，https://www.prrism.com/dpc/setumei_20180406.pdf，2019 年 12 月 10 日閲覧

8）田村一樹ら，評点付きレビュー文書を対象としたトピックモデルの構築に関する検討，情報処理学会論文誌，Vol. 56，No. 3，pp. 1013-1027，2015

9）David M. Blei, Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, pp. 993-1022, 2003

10）厚生労働省，平成 30 年度医療の質の評価・公表等推進事業公募要領　共通指標セット（参考資料 1），https://www.mhlw.go.jp/file/06-Seisakujouhou-10800000-Iseikyoku/0000166398_2.pdf,

2019 年 12 月 10 日閲覧
11）厚生労働省，事業場における治療と仕事の両立支援のためのガイドライン，
https://www.mhlw.go.jp/content/11200000/000490701.pdf，2019 年 12 月 10 日閲覧
12）国立がん研究センター，がん統計　最新がん統計，https://ganjoho.jp/reg_stat/statistics/stat/summary.html，2019 年 12 月 10 日閲覧