

トピックモデルを用いた病床機能報告データの分析

田中 昌昭^{*1}

^{*1} 川崎医療福祉大学 医療情報学科

Analysis of Clinical Function Report Data Using Topic Model

Masaaki Tanaka^{*1}

^{*1} Dept. of Healthcare Informatics, Kawasaki University of Medical Welfare

抄録：医療資源の有効活用と医療の最適化を目指して医療機能の明確化と機能分化が進められている。病床機能報告制度では診療実態を表すデータとともに病床機能の報告を義務付け、医療機関の自主的な現状把握と将来構想を促している。本研究では自然言語処理分野で培われたトピックモデルを用いて公表されている病床機能報告データを分析し、病床機能の特徴づける診療行為のパターンの抽出を試みた。その結果、医療機能を構成する機能単位をトピックとして抽出することができ、病床機能の定量的分析の可能性を示すことができた。

キーワード 病床機能報告, 医療機能分化, トピックモデル, 定量的分析, 地域医療構想

1. はじめに

高齢者人口が増える中、限りある医療資源を有効に活用するには地域において医療機関が担う医療機能を明確化し、機能分化を進めることが求められる。政府は医療機関の機能分化に向けての自主的な取り組みを促すために病床機能報告制度を設け、保有する病床が担う医療機能の現状をデータとともに報告することを義務付けた。収集されたデータは医療機関の実名とともに公表され、他の医療機関と比較が可能になった。

病床機能は高度急性期、急性期、回復期、慢性期の4つに区分されるが、それらの間に明確な判断基準がないため、報告されたデータは必ずしも実態に即しているとは言えない。そこで本研究ではデータ自体から4機能の特徴的なパターンを抽出して病床機能を識別する定量的な基準が得られないか検討することを目的とした。

2. 方法

1) 材料

平成29年度病床機能報告公表データ[1]の病棟票から「2017(平成29)年7月1日時点の機能」(以降、病床機能と記す)と報告様式2の「3. 幅広い手術の実施状況」から「12. 医科歯科の連携状況」までの112項目(以降、これを診療行為と記す)を取り出して分析対象データとした。

2) トピックモデルによる病棟票の分析

トピックモデル[2]は文書に隠されたトピックを推

定して文書分類や文書検索に利用する自然言語処理分野で培われた技術である。トピックモデルでは、文書はいくつかのトピックをある確率で含み(これをトピック分布と呼ぶ)、各々のトピックからはトピックごとに決まったある確率で単語が出現するとみなす。ここで、文書を病棟に、単語を診療行為に、そして単語の出現頻度を診療行為の算定回数に置き換えれば病棟のトピック分布及びトピックの診療行為分布を求めることができる。トピックモデルはPython3.6.9上でgensim.ldamodelを利用し、トピック数は36としてモデルを構築した。

3) 病棟の類似度の計算とそれに基づく分類

得られた病棟のトピック分布を用いて任意の病棟間の類似度を求めた。その際、類似度の尺度には確率分布の差異を表す Jensen-Shannon divergence を用いた。

次に、病棟票の診療行為を特徴量、病床機能を教師データとして k 近傍法を用いて病棟を分類した。その際、近傍の定義に用いる尺度として Jensen-Shannon divergence を用い、 $k = 20$ とした。また、通常のカテゴリ分類のようにデータを学習データと検証データに分けず、全データを使って学習した分類器を用いて同じ全データの分類を行った。

4) 病床機能を代表するトピック分布の算出

当該病床機能が割り当てられている全ての病棟のトピック分布との類似度の総和が最小になるようなトピック分布を求め、それをその病床機能を

代表するトピック分布と定義した。

3. 結果

1) 病棟のトピック分布

Fig.1 に得られた病棟のトピック分布をヒートマップで示す。横軸はトピック、縦軸は病棟で、病棟は「1. はじめに」に書いた病床 4 機能の順に並べ替えてある。

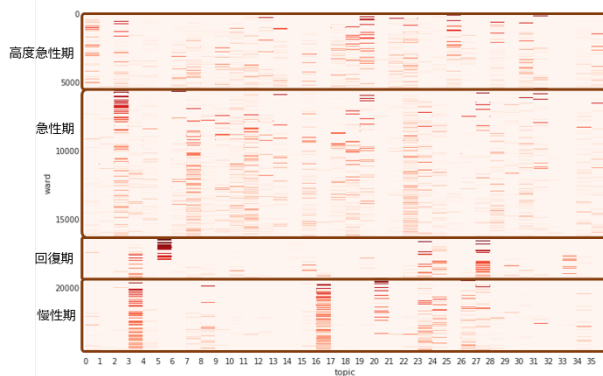


Fig.1 病棟のトピック分布

高度急性期病棟で顕著なトピック 0 (Fig1 の左端) には、手術総数、人工呼吸 (5 時間を超えた場合)、救命のための気管内挿管、全身麻酔の手術総数など高度急性期に特徴的な診療行為が高確率で含まれていた。

2) 病棟の分類結果

Table 1 に分類結果を示す。表側は報告された病床機能、表頭は k 近傍法の予測である。

Table 1 病棟の分類結果 (混同行列)

	高度急性期	急性期	回復期	慢性期	合計	再現率
高度急性期	3360	2017	8	44	5429	0.619
急性期	1068	9481	293	181	11023	0.860
回復期	11	218	2380	229	2838	0.839
慢性期	25	147	109	5145	5426	0.948
合計	4464	11863	2790	5599	24716	
適合率	0.753	0.799	0.853	0.919		0.824

正確度は 0.824 であるが、高度急性期の多くが急性期に誤分類され再現率は 0.619 であった。

3) 病床機能を代表するトピック分布

高度急性期、急性期を代表するトピック分布は、多彩なトピックから構成されている点と互いに共通するトピックが多い点で類似していた。一方、回復期と慢性期の場合は、構成するトピック数が少なく、それぞれの病床機能に対応した特定のトピ

ックが支配的になっている点で高度急性期や急性期とは異なっていた。

4. 考察

トピックモデルによって病床機能の特徴をトピックの含有パターンとして捉えることができた (Fig.1)。個々のトピックは診療行為の組み合わせからなり、各々の診療行為が含まれる割合がトピックごとに決まっている。これを、医療機能を構成する機能単位と考えれば病床機能はこの機能単位をどのような割合で含んでいるかで識別できる。これは、ある病棟がどの病床機能を有しているかを判断する際に定量的な基準を与える可能性を示唆している。

病棟の分類では学習データと検証データに同じデータを使っているにもかかわらず、正確度は 8 割強しかなかった。この原因は学習アルゴリズムが教師データ、あるいはその両方にある。いくら学習アルゴリズムが優れていても教師データが間違っていれば高い分類性能は得られない。今回の結果はまさにこれを示すもので、誤分類された病棟に本当に適切な病床機能が付与されているか検討の余地がある。Table 1 は高度急性期と急性期の間で誤分類が多いことを示しているが、これは Fig.1 や病床機能を代表するトピック分布が示しているように高度急性期と急性期の特徴が非常に似通っているからだと考えられる。

5. 結語

病床機能報告データにトピックモデルを適用することによって病床機能の定量的分析の可能性が示された。病床機能報告の定量基準分析として各都道府県から具体的な指標が提案されつつある[3]。本研究が定量基準構築の一助となれば幸いである。

参考文献

- [1] 厚生労働省: 病床機能報告, [https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000055891.html (cited 2020-Jan-31)].
- [2] 佐藤一誠: トピックモデルによる統計的潜在意味解析. コロナ社, 東京 2015.
- [3] 埼玉県: 定量基準分析結果, 2018. [https://www.pref.saitama.lg.jp/a0701/documents/bunseki_h29.pdf (cited 2020-Jan-31)]