一般口演A

[KA3] 一般口演 A

初診時記録及びオーダ情報に基づく病名登録率評価

2018年6月23日(土) 10:05 ~ 10:45 第1会場 (2階・メインホール)

[KA3] 初診時記録及びオーダ情報に基づく病名登録率評価

畠山 豊（高知大学医学部附属医学情報センター）

初診時記録及びオーダー情報に基づく病名登録率評価

島山 豊, 永田 桂太郎, 渡部 輝明, 奥原 義保

高知大学 医学部附属医学情報センター

Analysis of disease name registration based on records at first visit and order information

Yutaka Hatakeyama, Keitaro Nagata, Teruaki Watabe, Yoshiyasu Okuhara

Center of Medical Information Science, Kochi Medical School, Kochi

University

抄録: 大規模データ解析の入力として類似した背景を持つ患者群抽出可能性を評価するため、初診時記録及びオーダー情報によって分割された患者集団における各病名の登録率の評価を行った。初診時記録に対して潜在トピックモデルによって計算された各トピックに対する確率及びオーダーの有無による距離に基づき、それぞれ階層化クラスタリングを実施し、細分化されたクラスタ毎の登録率を取得した。高知大学医学部附属病院総合診療部受診の543人のデータに適用した結果、処方歴がある患者数5人以上のクラスタでの最大病名登録率は77%となった。特徴的な記述及び処方歴の集団では類似した背景を持つ患者群が抽出可能なことを示した。

キーワード 潜在トピックモデル、テキストマイニング、階層化クラスタリング

1. はじめに

病院情報システムデータによる大規模解析研究を実施する際、病名登録情報に基づいて対象患者を抽出する場合がある。この集団において検査値分布のバラつきなどの患者背景が揃わない問題があり、解析研究の精度を上げるためには様々な患者データを利用してより精度の良い同一背景を持つ対象データ抽出が重要である。

本論文では、類似した症状の記載及びオーダーがある同一の登録病名がある患者群を類似背景患者群として考える。高知大学医学部附属病院の総合診療部における初診時データに対し、初診時記録及びオーダー情報で特徴づけられた患者集団における病名を解析することで、類似背景を持つ患者群の抽出可能性評価を目的とする。

初診時記録及びオーダー情報によるそれぞれのクラスタリング結果に基づき患者集団を分割する。統合して分割した群ごとに病名を集計し、それぞれの登録率を評価する。

2. 方法

1) 初診時記録に基づく分類

テキスト情報を潜在トピックモデルによって特徴量を評価し、Jensen-Shannon ダイバージェンスによる階層化クラスタリングで患者を分割する。

形態素解析を実施し、動詞、形容詞、及び一

般名詞のみを対象とする。出現頻度が10未満の単語は除外し、抽出後の単語数が5未満の記録を除外した後、トピック数20で潜在トピックモデル[1]による解析を行う。出力としてテキスト毎の各トピックに対する確率 θ 及びトピック毎の各単語に対する確率 ϕ が得られる。

記録間のJensen-Shannon ダイバージェンスを確率 θ に基づき計算し、Ward法による階層化クラスタリングを実施する。出力結果から患者群を14クラスタに分割する。

2007年から2017年までの47人の医師が作成した543件の初診時記録データ(フリーテキスト)に対しComeJisyoV5辞書を用いNMecabをC#上で、階層化クラスタリングをR上で実施する。

2) オーダー情報に基づく分類

初診時にオーダーされた処方及び画像検査オーダーデータについて、医薬品の薬効分類情報ごとに処方の「あり」「なし」とした入力データに変換する。入力データのマンハッタン距離に基づくWard法による階層化クラスタリングを実施する。出力結果から患者群を12クラスタに分割する。

3) 病名登録率による評価

1)及び2)で作成したクラスタ情報を組み合わせた患者分割を行い、ICD10コード(3桁)登録情報を集計し分割集団ごとの登録率を取得する。

Table.1 初診時記録に基づく
クラスタリング結果

初診時 クラス	人数	トピック	Rank1	Rank2	Rank3	Rank4	Rank5
1	20	12	下痢	℃	腹痛	食べる	嘔吐
2	46	11	いる	こと	できる	本人	られる
3	64	5	仕事	こと	する	ストレス	不安
4	75	13	舌	便秘	腹部	脈	食欲
5	36	10	痛み	痛い	右	前	整形
6	64	3	咳	℃	痰	発熱	処方
7	21	18	人	いる	思う	歳	子供
8	56	6	する	内科	上	程度	認める
9	31	19	頭痛	めまい	感じ	肩こり	ない
10	18	15	いる	とき	する	学校	食べる
11	37	14	受診	こと	紹介	相談	希望
12	18	4	こと	処方	喫煙	飲酒	特記
13	28	20	受ける	希望	来院	not	診察
14	29	2	圧痛	可能性	貧血	出現	受診

3. 結果

初診時記録に基づき分類したクラスタでの患者数をクラスタ毎に集計し、最も多く選択されたトピック及び、そのトピックにおける確率 ϕ の上位 5 単語を示す。そのため、Table1 は各クラスタで使われやすい単語を示す。

オーダに基づく分類では、処方オーダがない患者のクラスタ(オーダクラス番号 2)が最も人数が多く(226 人)、漢方製剤を主に処方しているクラスタ(オーダクラス番号 3)が、2 番目に多い(130 人)結果となった。

初診時記録及びオーダに基づく結果から 94 クラスタに分割した。患者数の多いクラスタを Table2 に示す。各クラスタで最も登録患者数が高い ICD10 コードを示す。病名登録率はクラスタ患者数に対する登録患者数の比となる。初診時クラス番号 1 では症状の記載で登録病名が決定されていることを示している。初診時クラス番号 9 及び 12 では症状の記載だけでは登録病名が推定できず、処方の有無で病名が確定する状況を示している。5 人以上の患者数をもつクラスタ数が 28 で、患者数の 78%を占める。漢方製剤以外を主に処方する患者数 5 人以上のクラスタでの最大病名登録率は 77%となった。

4. 考察

初診時記録及びオーダデータから類似背景をもつ患者集団を抽出可能な結果が示された。

初診時記録を詳細に記載している総合診療部データを入力として用いることでテキストデータによる分割が良好な結果が得られ、高頻出の単語がトピック毎に異なる結果になったと考える。他の

Table.2 初診時記録及び
オーダ情報に基づくクラスタリング結果

初診時クラス	オーダクラス	患者数	ICD10	ICD10登録患者数
1	1	13	A09	10
1	6	4	A09	3
2	2	40	I10	3
3	3	18	F48	4
3	2	17	F41	5
4	3	57	R68	11
5	2	18	M54	6
5	3	9	M54	3
6	3	15	J00	6
6	8	14	J06	6
7	4	9	F32	9
7	2	5	F32	2
8	2	30	I10	3
9	2	13	R42	8
9	3	8	R51	4
10	2	10	R63	2
11	2	19	M47	2
12	2	9	I78	5
12	7	7	A37	7
13	2	20	R10	3
14	2	23	M62	2

診療科でも適用可能だが、対象疾患が限定的なため、単語のバラつきが小さい可能性がある。

2 つのクラスタリング結果に基づき細分化したクラスタを生成するが、実際に患者が存在するクラスタ数は約半数であり、ある程度の患者数があるクラスタはさらに少ない。この結果はテキスト情報とオーダ情報は関連していることを示し適切に特徴抽出を行った結果である。また、初診時クラスタ 12 の細分化結果が示すように、登録病名の絞り込みにも寄与していると考えられる。

本研究の限界は、処方情報が限定できない患者に対しては適切な患者背景情報が得られないことである。つまり、処方されていない患者群では十分な特徴情報が得られないため、病名登録率が低い傾向を示す。逆に漢方製剤を処方されている患者群では薬剤の適用症例範囲が広いと、処方情報が患者状態識別の情報量としては低くなる。そのため、処方情報がない群と同様に病名登録率が低い結果となっている。

5. 結語

初診時記録及びオーダ情報の特徴量に基づき分類した患者群での病名登録率を評価した結果、特徴的な記述及び処方の履歴がある集団では同一の病名を登録している傾向を示し、類似した背景を持つ患者群が抽出可能なことを示した。

参考文献

- [1] Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022. (2003)