6ZL-03

電子医療記録を用いた Latent Dirichlet Allocation による疾患共起トピックと個別疾患ごとのコスト期待値算出

田村夏生† 森雅也† 上村博輝†† 野中尋史†
長岡技術科学大学† 新潟大学††

1. はじめに

近年, 超高齢化や生活習慣病の増加により医療費が高騰している．そのため，健全な医療保険財政のための医療制度の改革や見直しが課題となっている．健康保険組合や行政などが持つ特定健診データや電子医療記録（レセプトデータ）を分析することで，高コスト疾患の把握や共起しやすい疾患群の特定が可能となる．これは効果的な疾病予防対策や将来の医療費の予測につながり，安定的な保険財政の運営に効果的である．

一方で，レセプトデータ分析は，患者の疾患とコストが直接対応していないため，個別疾患の正確なコストが把握できない問題が存在する．そのため，疾患のコスト計算を行うために疾患群として疾患をグループ化する必要がある． レセプトデータ分析の先行研究として，諸外国ではレセプトデータを記号列としてみなし, BERT に基づいて将来の疾患予測などを行う手法が多数開発されている[1] [2][3]．しかし，現在国内ではこのような研究はほとんど行われていない．

本研究のデータでは, BERT 系の手法を用いるにはレセプトデータ数が少なく，また海外の先行研究と傷病コードが異なるため転移学習も困難である．そこで，本研究では BERT 系の手法と比べてデータ数が少量でも学習しやすい階層的ベイズ手法である Latent Dirichlet Allocation（LDA）を採用した．LDA により一カ月分のレセプトデータをトピック（共起しやすい疾患群）に分類し，トピックごとのコストを算出することで，個別の疾患の点数を計算した. また, 分類されたトピックの分析を行った．

2. 実験概要

分析の対象とするデータは，新潟県妙高市の平成 29 年 3 月のレセプトデータ 5072 件である．前処理として，レセプトデータから傷病コードと合計点数を抽出し，患者ごとにまとめた．次に社会保険診療報酬支払基金の傷病名マスターを用いて，傷病コードを傷病名に変換した．

分析には, Python ライブラリの 1 つである GENSIM の LDA パッケージを用いた．ストップワードとして「未コード化傷病名」を指定し，すべて除外した．

LDA を用いた分析で重要な値は，トピックの選択確率を得るためのパラメータ α とトピック数である．デフォルトでは， α は $1/\alpha$ の値を適用するようになっている．今回は共起しやすい疾患が同じトピックに集まってほしいので，トピック数がより少ないことが望ましい．そこで， α の値を変化させながら平均トピック数を算出し，最も平均トピック数が少なかった $\alpha = 0.01$ を採用した．また，LDA を用いた分析では抽出するトピック数を事前に決定する必要がある．一般的にトピック数の決定には，確率モデルの精度を表す perplexity とトピックの品質を表す coherence が用いられる．本研究では，トピック数の値を変化させながら perplexity と coherence を算出し，トピック数を 4 と決定した．

疾患ごとの点数は，各トピックの上位 10 件の割合にトピックの点数をかけたものを足し合わせた値である．複数のトピックに同じ疾患が出てきた場合，重複する疾患はコストを足し合わせた．

3. 結果と考察

実行結果を pyLDAvis という LDA の結果をインタラクティブに出力するためのライブラリを用いて可視化した．結果を図 1 に示す．図 1 は各トピックを計量多次元尺度法(MMDS)で 2 次元に圧縮し, 配置したものを表示している．MMDS は，定量的数値データを対象に個体の類似度を計算し，最適な配置を算出するものである．配置された円の大きさは，各トピックに属するドキュメント(レセプトデータ)の合計を表しており，円同士の距離はトピック間の距離を表している．後述する各トピックの疾患を踏まえると，縦軸は性別を表しており横軸は年齢を表していると考えられる．理由としては，縦軸は上側では男性要素が強くなり，下側では性別の要素が弱くなっていること，横軸は右側では老視など罹患する年齢が高い疾患を持つトピックがあり，左側にはインフルエンザなど年齢にかかわらず罹患する疾患があることが挙げられる．

各トピックの疾患とトピックごとのコストを表 1 に示す．トピックごとの疾患はトピック内の推定用語頻度が全体の用語頻度の半分以上あるものを選んだ．また，各トピックの総コストは，患者ごとのレセプトがそれぞれ

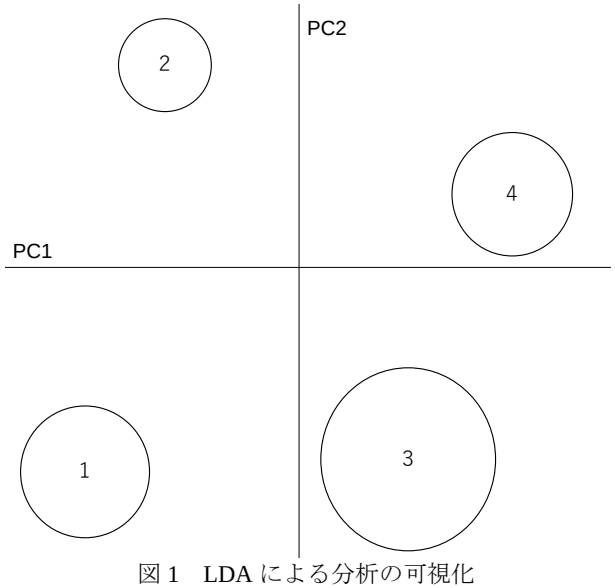

図 1 LDA による分析の可視化

Calculation of disease co-occurrence topics and cost expectations for each individual disease by Latent Dirichlet Allocation (LDA) using electronic health records (receipt data)
†Natsuo Tamura, †Masaya Mori, ††Hiroteru Kamimura and †Hirohumi Nonaka
†Nagaoka University of Technology
††Niigata University

表 1　各トピックの疾患とコスト

	トピック 1	トピック 2	トピック 3	トピック 4
傷病名	1.変形性膝関節症 2.変形性腰椎症 3.肩関節周囲炎 4.急性気管支炎 5.インフルエンザ 6.腰部脊柱管狭窄症 7.疼痛 8.インフルエンザ A 型 9.急性上気道炎 10.急性咽頭炎	1.前立腺がん 2.神経因性膀胱 3.過活動膀胱 4.下腹痛 5.膀胱がん 6.アルコール性肝炎	1.高血圧症 2.高脂血症 3.不眠症 4.糖尿病(2 型糖尿病) 5.高コレステロール血症 6.慢性胃炎 7.閉経後骨粗鬆症 8.脂肪肝 9.不安神経症 10.心肥大 11.うつ病 12.統合失調症	1.老視 2.遠視性乱視 3.加齢性白内障 4.慢性結膜炎 5.アレルギー性結膜炎 6.眼内レンズ挿入眼 7.近視性乱視 8.ドライアイ 9.糖尿病網膜症
コスト	491	393	779	468

表 2　疾患ごとのコスト

傷病名	コスト
高血圧症	171
高脂血症	98
胃潰瘍	92
便秘症	73
糖尿病	71
アレルギー性鼻炎	56
慢性胃炎	45
不眠症	33
変形性膝関節症	32
老視	27
変形性腰椎症	26
血尿	26
遠視性乱視	24
加齢性白内障	22
肩関節周囲炎	21
高コレステロール血症	21
前立腺肥大症	21
急性気管支炎	20
慢性結膜炎	19
前立腺がん	18
2 型糖尿病	18
アレルギー性結膜炎	18
インフルエンザ	17
骨粗しょう症	17
閉塞性動脈硬化症	16
高尿酸血症	15
内痔核	13

のトピックに対し，どの程度の確率で該当するか算出し，その確率にレセプトデータごとの合計点数を掛けた値をトピックごとに足し合わせた値とした．

表 1 を見ると，一番コストが高いのがトピック 3 であり，主に生活習慣病が分類されていることがわかる．次にコストが高いのがトピック 1 で，関節症とインフルエンザなどの風邪の疾患が分類されている．これは，分析に使用したレセプトが 3 月のものなので，風邪の疾患が多く共起したと考えられる．また，関節症と風邪の関係を示すような研究結果はなかった．トピック 1 より少し低いコストとなったのが，トピック 4 である．このトピックは眼に関する疾患が集まっている．一番低いコストとなったのがトピック 3 である．主に膀胱に関する疾患が集まっており，さらに一見関係ないようなアルコール性肝炎が含まれている．しかし Lao[4]らによると，男性の場合，アルコール摂取によって膀胱がんの発症確率が増加することが示されている．このことから，このトピックはアルコールを大量摂取した男性がかかる疾患が集まっている可能性がある．コストが一番低い理由としては，このトピックは全体の人数が少ないからではないかと考えられる．

最後に，疾患ごとのコストを表 2 に示す．特徴としては高血圧症や高脂血症などの生活習慣病に関する疾患が特に高コストであることが分かった．特に高血圧は複数のトピックで現れており，それぞれで高コストであったため特にコストが高くなっている．

4. おわりに

本研究では，妙高市のレセプトデータにおいて，疾患とコストの対応付けする方法の提案とトピックの分析を行った．今後の予定として，LDA の改良に関する検討（中華料理店過程など）や健康診断データと組み合わせて年齢や性別による影響を検討する．

参考文献

[1] Li, Y., Rao, S., Solares, J.R.A. *et al.* BEHRT: Transformer for Electronic Health Records. *Sci Rep* **10,** 7155 (2020).

[2] Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W. Jim Zheng, Kirk Roberts, Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review, Journal of Biomedical Informatics, Volume 115,2021,103671,ISSN 1532-0464.

[3] Pang, C., Jiang, X., Kalluri, K.S., Spotnitz, M., Chen, R., Perotte, A. & Natarajan, K.. (2021). CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. 158:239-260.

[4] Lao, Y., Li, X., He, L., Guan, X., Li, R., Wang, Y., Li, Y., Wang, Y., Li, X., Liu, S., & Dong, Z. (2021). Association Between Alcohol Consumption and Risk of Bladder Cancer: A Dose-Response Meta-Analysis of Prospective Cohort Studies. Frontiers in oncology, 11, 696676.