一般口演

一般口演18
医療データ分析６（機械学習・テキストマイニング１）

2018年11月25日(日) 09:00 〜 10:30 D会場 (4F 413+414)

[4-D-1-1] 非構造化テキストデータの活用による Phenotypingの精度向上

○山下 貴範[1], 伊豆倉 理江子[1], 野原 康伸[1], 野尻 千夏[2], 高田 敦史[1], 中島 直樹[1] （1.九州大学病院　メディカル・インフォメーションセンター, 2.株式会社　ケア・フォー）

【はじめに】平成30年度より本格稼働した MID-NET事業において、副作用調査としての正確な病態抽出は重要である。我々は、病院情報システム（ HIS）のデータから一定の精度を担保する Phenotyping手法の開発を進めている。伊豆倉ら（2018 JAMI春季）は、 HISの患者属性、医薬品、検査結果などの構造化データを用いて間質性肺炎（以下 IP）の Phenotyping手法を開発した。具体的には「初期ルール」で広めに抽出した症例からランダムサンプリングし、専門医レビューで真偽判定した後、目的変数を真症例、説明変数を構造化データとして、勾配ブースティング（ GBDT）と rpartによる AUC=0.798の「抽出ルール」を策定した。次に、初期ルール以外の真症例を CTレポートテキスト検索で抽出し、この真症例を構造化データ GBDTで予測したが、 AUCは0.582と低く追加抽出するルール策定ができなかった。本研究では、非構造化テキストデータからの真症例抽出ルール策定を目的とした。【方法】当院の全入外患者117,401名（2014年〜2015年）の CTレポートから、 IP症例を推定する8キーワードが含まれる5,141名を抽出した。それから100名をランダムサンプリングし、先行研究の初期ルールに含まれない84名を選定した。専門医レビューでの真症例は39名であった。84名の CTレポートテキストに対して形態素解析を行い、987の変数を抽出した。目的変数を真症例、説明変数を①先行研究＋テキスト、②テキストとして、 GBDTで予測した。【結果・考察】 AUCは①0.653、②0.732であり、②の方が改善された。今回の対象84名には IP診断が付いておらず、 IPに関連する処方や検査が少ないことで、②の方が良いモデルとなり AUCが改善したことが考えられる。また変数重要度から、①で特定できなかった IPに関連するワードが②では上位に現れた。非構造化テキストデータ解析による真症例の抽出精度を向上することができた。

©一般社団法人 日本医療情報学会

非構造化テキストのデータ活用による Phenotyping の精度向上

山下貴範[*1]、伊豆倉理江子[*1]、野原康伸[*1]、野尻千夏[*2]、高田敦史[*1]、神田橋忠[*1]、中島直樹[*1]
*1 九州大学病院 メディカル・インフォメーションセンター、*2 株式会社 ケア・フォー

Improvement of Phenotyping algorithm using Text-based Records

Takanori Yamashita[*1] , Rieko Izukura[*1] , Yasunobu Nohara[*1], Chinatsu Nojiri[*2], Atsushi Takada[*1], Tadashi Kandabashi[*1], Naoki Nakashima[*1]
*1 Medical Information Center, Kyushu University Hospital, *2 Care Four Inc.,

We have developed some phenotyping algorithms to ensure high performance from real word data. Izukura et al. (2018) developed phenotyping of interstitial pneumonia (IP) using machine learning from structured data (patient information, medicine, laboratory result and diagnosis) in the hospital information system. However, it could not make an algorithm from CT reports (unstructured text-based data) because AUC was low. In this research, we aimed to develop true IP case extraction an algorithm from unstructured text-based data. 5,141 patients including some keywords for predicting IP cases in CT report were extracted. Then, 100 patients were randomly sampled and 84 patients not included in the algorithm of the previous study were selected. In order to analyze the 84 CT reports , we constructed a search engine of the records. Morphological analysis was applied to the CT reports, 987 words were extracted and true case was predicted by GBDT, AUC was 0.732. We extracted 4 keywords as feature word from its score, and PPV is 0.548 and sensitivity is 0.872 when one of them is included. Phenotyping performance by unstructured text-based data analysis could be improved.

Keywords: RWD (Real World Data), Phenotyping, Interstitial pneumonia (IP), GBDT.

1. はじめに

近年、医療分野ではビッグデータ、リアルワールドデータ（以下、RWD）の活用により、これまでの臨床研究では分からなかった治療法や副作用検知、発症予測・予防、創薬の開発が期待されている。その RWD を蓄積する病院情報システム（以下、HIS）は、医事会計システムから発展しており、医事会計のための診療行為の医薬品、検査結果、病名などはコード化された構造化データとして蓄積されている。

一方、患者状態や診療行為の目的、その結果（アウトカム）はカルテやレポートにフリーテキストにて記載されている。本来は、「患者の症状→医師による診察、検査→検査結果→診断（病名）→治療（医薬品、手術など）→アウトカム）」が診療プロセスとして関連付けされるが、現在の HIS では、それぞれの診療行為のデータが独立して蓄積されている。そのデータから、何の病気で、どのような検査結果と医師の判断による治療を経て、どのようなアウトカムになったのかを正確に抽出することは困難な状況である。そのため RWD 活用から、正確な病気（phenotype）の患者群を一定のルールで再現性良く抽出する方法（phenotyping）の開発が重要である。

また、これらは平成 30 年度より本格稼働した MID-NET 事業において、副作用調査としての正確な病態抽出も必要であり、著者らは、HIS に蓄積されたデータから一定の精度を担保する Phenotyping 手法の開発を進めている。

2. 先行研究

島井ら（2014）[2)]は、CT レポートをランダムサンプリングし、専門医レビューによる間質性肺炎の診断と否定した症例について分類し、そのレポートの非構造化テキストデータと KL6、SP-D の数値データを用いた。テキストと数値データの陽性尤度比と陰性尤度比を算出し、それぞれ積算した得点の精度検証から間質性肺炎を特定している。

伊豆倉ら（2018）[1)]は、HIS の患者属性、医薬品、検査結果などの構造化データを用いた Phenotyping 手法を開発した（図 1）。具体的には、専門医やガイドラインを基に定義した「初期ルール」で広めに抽出した症例からランダムサンプリングし、専門医レビューで真偽判定した後、目的変数を真症例、説明変数を構造化データとして、勾配ブースティング（GBDT）と決定木（rpart）による AUC=0.798 の「抽出ルール 1」を策定した（図 1・STEP1）。次に、初期ルール以外の症例について、CT レポートから、間質性肺炎に関連するキーワードをテキスト検索で抽出し、専門医レビューで真偽判定した後、真症例を目的変数、STEP1 で利用した構造化データを説明変数として GBDT で予測したが、AUC は 0.582 と低く、追加抽出するルール策定ができなかった（図 1・STEP2）。しかし、STEP2 からも真症例が抽出されたことで、ALL possible case の感度（約 10%）が精緻化され、さら MID-NET にて適用可能なロジックであることを示唆している。。

本研究では、伊豆倉ら（2018）[1)]の課題となった STEP2 の「抽出ルール 2（図 1）」の策定を目的として、CT レポートの非構造化テキストデータを対象として、検索エンジンを構築し、機械学習を用いて解析した。

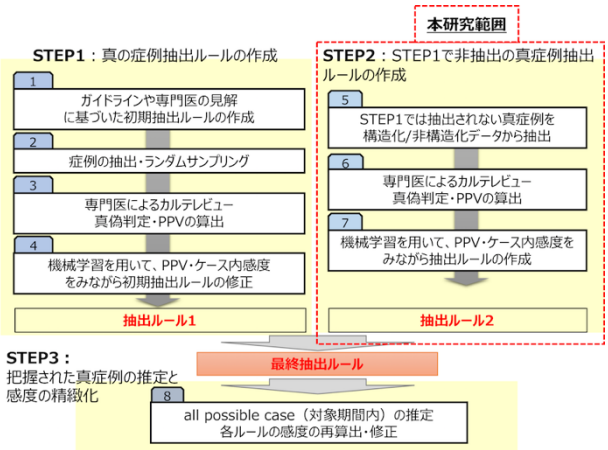

図１ Phenotyping の開発工程（伊豆倉ら(2018)より引用）

3. データ・方法

伊豆倉ら(2018)[1]と同じく、九州大学病院の2014年～2015年の入外患者 117,401 名の内、CT レポートから間質性肺炎を推定する 8 つのキーワード「間質、網状、すりガラス、スリガラス、蜂巣肺、器質化肺炎、間質性肺炎、interstitial pneumonia (IP) 」が含まれる 5,141 名を抽出した(STEP2-5)。それから 100 名をランダムサンプリングし、伊豆倉ら(2018)[1]の「初期ルール」に含まれない 84 名を選定し、専門医レビューによる真偽判定を実施した(STEP2-6)。

次に、84 名の CT レポートの非構造化テキストデータに対して、患者(文書)単位の検索エンジンを構築した。検索エンジンは NII で公開されている GETA[3]を利用した。84 文書から、987 の単語が抽出された。目的変数を真症例、説明変数を①伊豆倉ら(2018)[1]の構造化数値データ(医薬品、検査値、病名、患者属性)＋CTレポート単語、②患者属性＋CTレポート単語として、それぞれ GBDT で予測した。

さらに、GBDT の結果から特徴量の高い単語について注目し、それぞれ PPV と感度を推定した。また、検索エンジンによる SMART 類似度[4]の高い単語を用いて、真症例に対する単語の組合せにより PPV と感度を推定した(STEP2-7)。

4. 結果

CT レポートに対する専門医レビューでの真症例は 39 名であり(kappa 係数:k=0.61)、陽性的中度(PPV)は 0.464 であった。これから対象期間内の STEP2 で推定される真症例は、2.387 であった。

GBDT による結果の AUC は、①構造化数値データ(医薬品や検査結果)と CT レポート単語を組合せた場合が 0.653(図 2)、②CTレポート単語のみの場合が 0.732 であり(図 3)、②の方が改善された。②で抽出された特徴量が上位で、間質性肺炎に関連する単語について、それぞれ PPV と感度を算出した(表 1)。「網状、両肺」については、PPV と感度共に比較的高い値であった。キーワードの一つである「器質化肺炎」は、低い値となった。

検索エンジンにて、真症例対する SMART 類似度の高い単語として「網状、両肺、胸膜下、複数」が抽出された。その 4 つの単語の OR 条件では 62 名が対象となり、PPV(0.548)と感度(0.872)となった(表 2)。AND 条件では、3 名しか対象ではなかったが、全て真症例であった。これら 4 つの単語の前後の単語や文に注目すると、網状は「網状影」、両肺は「両肺上葉、両肺下葉背側、両肺底部」、胸膜下は「両肺胸膜下、肺下葉胸膜下、胸膜下優位に」、複数は「複数の腫大リンパ節、複数の小結節、複数の胆石」が見られた。

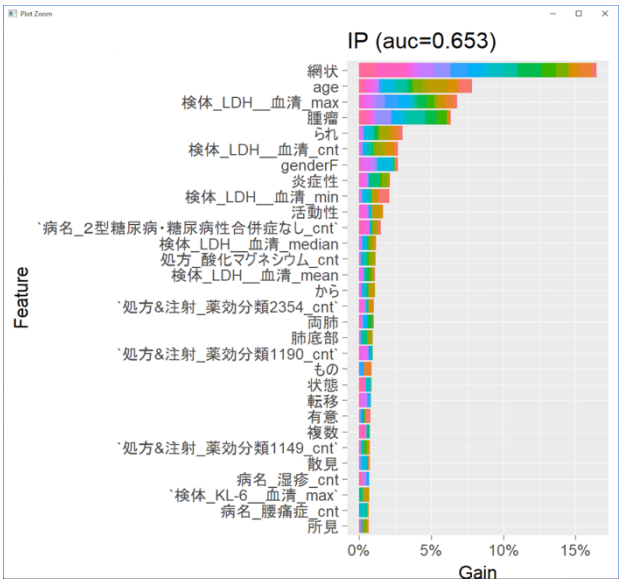

図 2 GBDT による解析結果①

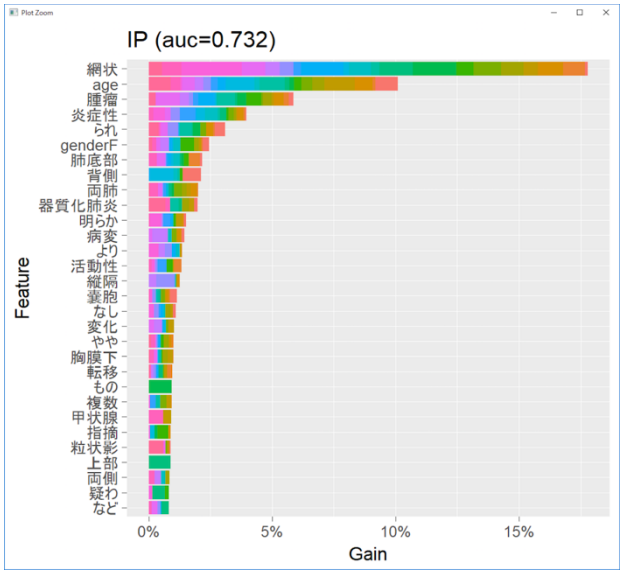

図 3 GBDT による解析結果②

5. 考察

本研究では、比較的解析が困難である非構造化テキストデータを対象にして、Phenotyping の開発と改善を試みた。伊豆倉ら(2018)[1]では、いわゆる構造化データを対象にした Phenotyping である。それでは抽出されない真の症例について非構造化テキストデータで抽出、検証できることを示すことができ、「網状、両肺、胸膜下、複数」の OR 条件を抽出ルール 2 を策定した。1 章で述べた現在の HIS の特性上、診療プロセスに沿って診療データが蓄積されていないため、真の症例を確実に抽出することが困難である。その中で、この Phenotyping 手法が有効であることを示した。最終抽出ルール(図 1)については、現在検討中である。

また、今回のランダムサンプリング対象 84 名には間質性肺

炎の診断が付いておらず、間質性肺炎に関連する処方や検査が少ないことで、②の方が良いモデルとなり AUC が改善したことが考えられる。また、変数重要度や類似度を考慮して、非構造化テキストデータ解析による真症例の抽出精度を向上することができた。

5. 結語

間質性肺炎の Phenotyping アルゴリズムについて、先行研究で課題となっていた非構造化テキストデータからの抽出ルールを策定した。

参考文献

1) 伊豆倉理江子，野原康伸，山下貴範 ら．病院情報システムの構造化データから病態を抽出する Phenotyping 手法の開発．第 22 回日本医療情報学会春季学術大会，2018：126-127.
2) 島井良重，武田理宏，真鍋史朗，寺本圭，三原直樹，松村泰志．電子カルテデータを用いた間質性肺炎の推測 －薬剤有害事象発生の把握を目的として－．第 34 回医療情報学連合大会論文集，2014：34，558-591.
3) http://geta.ex.nii.ac.jp/geta.html
4) 西岡真吾．汎用連想計算エンジン GETA．コンピュータソフトウェア，2009：26(4)，87-106.

表 1　GBDT による特徴語の PPV と感度

単語	PPV	感度
網状	0.742	0.590
腫瘤	0.310	0.231
炎症性	0.240	0.154
肺底部	0.714	0.256
背側	0.600	0.231
両肺	0.543	0.487
器質化肺炎	0.091	0.026
縦隔	0.385	0.385
嚢胞	0.591	0.333
胸膜下	0.611	0.282

表 2　SMART 類似度による 4 単語 OR 条件の PPV と感度

単語	PPV	感度
網状、両肺、胸膜下、複数	0.548	0.872