一般口演 | 医療データ解析
一般口演16
医療データ解析
2019年11月23日(土) 14:40 ～ 16:40 B会場 (国際会議場 2階コンベンションホールB)

[3-B-2-01] 電子カルテに記載されたテキストを対象とした機械学習による日単位での嘔気嘔吐症状の有無判定

○國方 淳[1]、十河 智昭[2]、間島 行則[1]、谷川 雅俊[2]、横井 英人[1,2]（1. 香川大学医学部附属病院 臨床研究支援センター, 2. 香川大学医学部附属病院 医療情報部）

キーワード：Machine Learning, Natural Language Processing, Text Mining, Electronic Health Record

【1. 背景と目的】電子カルテには症状についての情報が数多く記載されているが，その多くは構造化されたデータとして記録されておらず，症状の有無を機械的に抽出する方法は確立されていない．そこで，今回我々は電子カルテに記載されたテキストを対象とし，機械学習の手法を用いて日単位での嘔気嘔吐症状の有無判定を試みた．

【2. 方法】当院の2015年10月の入院患者の電子カルテのデータより，医師記載，看護師記載，リハビリ等の実施記録，救急診療記録をテキストデータとして抽出した．10月1～24日のデータを学習データ，25～31日のデータをテストデータとし，患者ごとに1日単位で嘔気嘔吐症状のラベリングを行い，前処理として文の絞り込み，形態素解析， Bag of Wordsもしくは分散表現への変換等を行った上でナイーブベイズ，ロジスティック回帰，サポートベクトルマシン，勾配ブースティング木，ニューラルネットワークのアルゴリズムを用いて学習と評価を行った．

【3. 結果】学習データ10805日分のうち嘔気嘔吐ありとラベリングされたのは651日，テストデータ2951日分のうち嘔気嘔吐ありとラベリングされたのは179日であった．今回使用したアルゴリズムの中では勾配ブースティング木の結果が最も良い性能を示し，テストデータにおいて Precision 0.87， Recall 0.73， F1 score 0.79であった．

【4. 考察と結論】機械学習による症状の有無判定の性能はテキストの記載様式による影響が大きく，テキストの絞り込みの方法が性能を大きく左右した．また，嘔気嘔吐は症状がない場合にも「嘔気はない」といった記載がなされることが多いが， word 5-gramを入力データとして使用することである程度の事実性の判定が可能となった．より大規模なデータでは，ニューラルネットワークが有望な選択肢となることが予想される．

電子カルテに記載されたテキストを対象とした機械学習による日単位での嘔気嘔吐症状の有無判定

國方 淳[*1]、十河 智昭[*2]、間島 行則[*1]、谷川 雅俊[*2]、横井 英人[*1,2]
*1 香川大学医学部附属病院 臨床研究支援センター
*2 香川大学医学部附属病院 医療情報部

Daily symptom detection for nausea and vomiting from texts written in electronic health records by machine learning.

Jun Kunikata[*1], Tomoaki Sogo[*2], Yukinori Mashima[*1], Masatoshi Tanigawa[*2], Hideto Yokoi[*1,2]
*1 Clinical Research Support Center, Kagawa University Hospital
*2 Department of Medical Informatics, Kagawa University Hospital

Although electronic health records contain a lot of information about patient's symptoms, most of them are not structured, and thus the way to detect symptoms automatically is not established. Therefore, we attempted to detect nausea and vomiting on a daily basis from texts written in electronic health records by machine learning.

We extracted medical records, nursing records, implementation records (e.g. rehabilitation), and emergency medical records recorded in October 2015 from our hospital's medical health record system. Among them we assigned data from October 1st to 24th to the training data and that from 25th to 31st to the test data. We labeled them presence of nausea or vomiting by patient and by day. After preprocessing, we trained models including naive Bayes, logistic regression, support vector machine, gradient boosting decision tree, and neural network with the train data and evaluated them with the test data.

The train data contains 10,805 days data in which 651 days were labeled with having nausea or vomiting and the test data contains 2,951 days data in which 179 days were labeled with having nausea or vomiting. Logistic regression, gradient boosting decision tree, and support vector machine showed similar performance with F1-score around 0.80 and precision-recall AUC around 0.89.

The performance of the models heavily depends on the format of texts and how texts were selected. Negative expressions about symptoms often causes problems like false positive, but word N-gram seems to be helpful for improving the performance to a certain extent.

Keywords: Machine Learning, Natural Language Processing, Text Mining, Electronic Health Record

1. 緒論

近年、多くの医療機関で電子カルテが導入されており、診療録をはじめとする多くの診療情報が電子的に保存される状況となっている。電子的に保存された診療情報のうち、病名や処方などの情報は定型的なフォーマットに従って記録されているためコンピュータを用いた大規模な分析を比較的容易に行うことができる。一例として、厚生労働省と独立行政法人医薬品医療機器総合機構（PMDA）が構築している医療情報データベース「MID-NET」では 400 万人を超える規模の医療情報を収集・分析することが可能とされている[1]。

一方で、患者の症状の多くは自由記載の形で記録されているため、症状の有無を機械的に判定するのは困難なのが現状である。しかし、症状の有無を機械的に判定することが可能であれば、薬剤による副作用の発現頻度や治療・看護の違いによる症状の発現頻度などの分析を大規模に行うことが可能になると考えられる。

症状をテキストデータから抽出するための試みとしては、症状や病名に関連する語を広く抽出したデータベースである万病辞書[2]や、患者が記述した文章から症状表現を抽出するソフトウェアである AEX (Adverse Effect eXtractor)[3]などが作成されている。これらは症状の表現を網羅的に扱うことを目的としているが、より限局的なアプローチとして特定の症状の有無をターゲットとする方法も考えられる。このアプローチでは個々の症状について判定のアルゴリズムを作成する必要があるが、それぞれの症状に特異的な特徴をアルゴリズムに組み込むことができるため精度面では有利になる。そこで、今回我々は嘔気・嘔吐の症状に焦点を当てて、入院患者について日単位での嘔気・嘔吐症状の有無を機械学習によって判定できるかどうか検討を行った。

2. 目的

電子カルテから抽出した入院患者の診療記録のテキストデータに対して自然言語処理および機械学習の手法を適用し、患者ごとに1日単位で嘔気・嘔吐の症状の有無を判定する。機械学習には複数のアルゴリズムを使用し、それぞれについて性能評価を行う。

3. 方法

本研究は香川大学医学部倫理委員会の承認（受付番号: 平成 28-065）の下で実施した。プログラム言語として Python 3.6.4 を使用し、Python の機械学習ライブラリとして scikit-learn 0.19.1、XGBoost 0.80、Keras 2.2.4 を用いた。また、形態素解析ソフトウェアとして MeCab 0.996 [4]を使用し、MeCab のシステム辞書には IPA 辞書を用いた。

3.1 テキストデータの抽出と分割

香川大学医学部附属病院の 2015 年 10 月の電子カルテデータより、入院患者を対象とする医師記載，看護師記載，リハビリ等の実施記録，救急診療記録をテキストデータとして抽出した。上記以外のデータ（テンプレートの記載や検温表の記載などを含む）は対象外とした。10 月 1 日から 24 日のデータ

を学習データ，25 日から 31 日のデータをテストデータとした。

3.2 ラベリング

評価者がテキストデータを読み、記載日に嘔気または嘔吐の症状があると判断された場合を「症状あり」、ないと判断された場合を「症状なし」とラベリングした。「昨日は嘔気があったが今日は改善している。」は「症状なし」とラベリングされるが、夜間の嘔気嘔吐症状の記載については日付が変わる前後いずれかに関わらず「症状あり」とした。

ラベリングの効率化のために、表 1 に示す嘔気嘔吐関連文字列で OR 検索を行い、これらの語句の周辺を読んで症状の有無を判断する方法を用いた。この方法の妥当性を確認するためにテストデータ 3 日分については検索を用いずに全てのテキストデータを目視して嘔気嘔吐症状の有無を確認したが、検索から漏れた嘔気嘔吐症状の記載は存在しなかった。

表 1　嘔気嘔吐関連文字列

悪心	おしん	嘔	吐	おうと	おうき
はいた	はいて	はき	はく	vom	naus
戻す	戻し	もどす	もどし	気分	きぶん
気持ち	きもち	ムカ	むか	胃部不快	えずく
えずき	えずい	えづく	えづき	えづい	

嘔気・嘔吐の症状についての記載のスクリーニングに用いる。そのため、頻度の高い誤変換や誤字も含めている。vom は vomit の、naus は nausea の一部である。

3.3 前処理

当院の電子カルテにおいて医師記載は「Problem List」「Subjective」「Objective」「Assessment」「Plan」の 5 つのセクションから構成されるが、Problem List には過去の情報が多く含まれるためこのセクションはデータから削除した（ただし、Problem List にその日の嘔気嘔吐の症状が記載されている場合、ラベリングとしては「症状あり」となる）。

嘔気嘔吐に関連する文字列を含まない文は判定への寄与度が低いと考えられるため、ラベリングで用いたのと同じ嘔気嘔吐関連文字列を含まない文を学習データおよびテストデータのテキストから除去した。その結果として文が残らなかった場合、学習データについては除外し、テストデータについては無条件で「症状なし」と判定することとした。

その他の前処理として全角文字と半角文字の統一、形態素解析ソフトウェア MeCab による単語への分割、一部の単語の表記揺れの統一、ストップワードの除去、文開始・文終了タグの付与を行った後に、word 5-gram までの語句のうち 3 回以上出現した語句を用いて bag of words を作成した。また、ニューラルネットワークへの入力用にランダムな単語分散表現への変換を行った。

3.4 学習と評価

機械学習のアルゴリズムとしてナイーブベイズ，ロジスティック回帰，サポートベクトルマシン，勾配ブースティング木，ニューラルネットワークのそれぞれについて学習と評価を行った。ニューラルネットワークを除くアルゴリズムのパラメータはグリッドサーチを用いて決定した。ニューラルネットワークについては全結合層のみを用いたモデルと 1 次元畳み込みニューラルネットワークについて検討を行った。

4. 結果

学習データは 10,805 日分のテキストデータとなり、そのうち 651 日が「症状あり」とラベリングされた。テストデータは 2,951 日分で、そのうち「症状あり」は 179 日であった。学習データのうち 2,702 日分は検証用データとし、残りの 8,103 日分から嘔気嘔吐関連文字列を含まないデータを除外した 3,408 日分のデータを実際の学習に使用した。前処理後の学習データの例を表 2 に示す。

表 2　学習データの例

1 日分のテキスト	症状
<BOS> 頭痛 嘔気 ない <EOS> <BOS> 嘔気 ない 食事 摂取 できる て いる <EOS>	なし
<BOS> 嘔吐 <EOS> <BOS> PET 検査 ため 注射 後 帰室 途中 に 少量 嘔吐 あり <EOS> <BOS> 嘔吐 後 嘔気 消失 <EOS>	あり
<BOS> フェルムカプセル 14 日 分 手渡す 本日 分 内服 確認 する <EOS>	なし

文頭に<BOS>タグ、文末に<EOS>タグを付与した上で、MeCab により単語への分割と原形への変換が行われている。3 つ目の例では「ムカ」の文字列により嘔気嘔吐と無関係な文が選択されているが、この時点で多少のノイズが入り込むことは大きな問題とはならない。

各アルゴリズムのテストデータに対する評価結果を表 6 に示す。「症状あり」の割合が小さい skewed data であるため ROC-AUC はいずれも高くなったことから Precision-Recall AUC（PR-AUC）を主要な評価項目とした。ロジスティック回帰、勾配ブースティング木、サポートベクトルマシンがいずれも PR-AUC 0.89 とほぼ同等の性能を示した。ナイーブベイズの ROC 曲線と PR 曲線を図 1 に、勾配ブースティング木の ROC 曲線と PR 曲線を図 2 に示す。

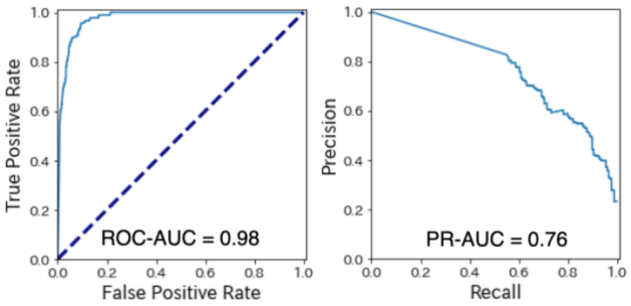

図 1 ナイーブベイズの ROC 曲線（左）と Precision-Recall 曲線（右）

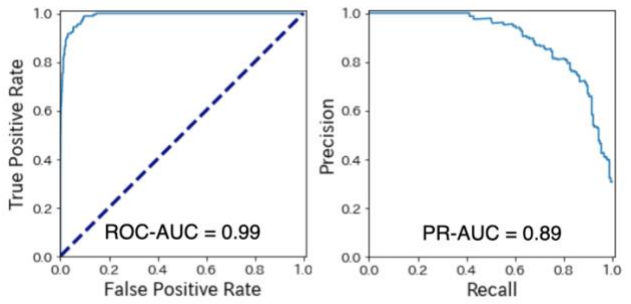

図 2 勾配ブースティング木の ROC 曲線（左）と Precision-Recall 曲線（右）

ニューラルネットワークについては、全結合層のみを用いたモデルにおいて単語分散表現の次元を 10 と小さくし、全結

合層もノード数 20 の1層のみというごく小さいモデルであっても容易に過学習を起こした。また、1 次元 CNN についても単語分散表現の次元を 32 と比較的小さく設定し、畳み込み層も2層のみの単純なモデルを用いたが同様に過学習となった。いずれの場合も過学習になる前に学習を打ち切ったモデルにおいてナイーブベイズよりはやや高い性能を示したものの、その他のアルゴリズムには及ばなかった。

5. 考察

一般的なセンチメント分析と異なり、特定の症状の有無を判定する場合はテキスト全体のごく一部のみが判定に大きく影響する。そのため、嘔気嘔吐に関連する文字列であらかじめテキストを絞り込む手法が有効であった。この絞り込みの段階では嘔気嘔吐に関係のない文の混入はそれほど問題とならないため、見逃しを防ぐために広めに拾うよう設定した。この手法はデータ量が比較的少ない場合には有効であるが、大量の学習データを扱える場合は逆効果となりうる。

電子カルテ上の記載は通常の文章と比べて不完全な文である場合が多い。例えば単に「肺炎」と記載があれば、多くの場合それは患者が肺炎に罹患している状態であることを示す。また、症状はしばしば「胃腸炎による下痢」のように体言止めで記載される。そのため、文頭と文末のタグは比較的大きな役割を果たすと考えられる。実際、表 3 に示すロジスティック回帰における係数や図 3 に示す勾配ブースティング木の変数重要度において文頭・文末タグの占める割合は比較的高い。

今回 word 5-gram を用いたのは、文頭・文末に加えて否定語や時期を表す語などの語順を考慮した判定が行われることを期待してのことである。否定語については一定の効果が認められていると思われるが、時期を表す語については表 4 の 1 例目の誤判定例に見られるようにあまりうまく判定できていない例が多かった。これは、時期を表す語が嘔気嘔吐を表す語から離れているケースが多いことに起因すると考えられる。bag of words でこの種の誤判定を防ぐためには、過去の事象を表す記載部位を前処理の段階で削除する等の対策が必要であろう。

本研究で用いたアルゴリズムのうち、一般的に性能が高いとされる勾配ブースティング木とサポートベクトルマシンとともに、比較的単純なアルゴリズムであるロジスティック回帰でも前二者と同等の性能が得られた。一方で、ニューラルネットワークではデータ量の不足から過学習を起こし、十分な学習ができなかった。大量のデータを利用できる場合は性能の向上が期待できるが、学習データのラベリング作業が課題となる。

本研究の限界としては、学習データとテストデータがともに 1 ヶ月間という狭い範囲内のものであるため、勤務している医師や看護師の個人的な記載の癖も学習している可能性が高い。学習後に実際に判定を行うのは未来のデータであることを想定してテストデータは 1 ヶ月の最後の 7 日間としたが、時系列的に連続したデータであるため同一の患者が学習データとテストデータに含まれることになる。これらの影響を考慮すると、ある程度の期間を置いて勤務者や患者が入れ替わったテストデータでの評価も必要である。また、理想的な Gold Standard として患者の1日ごとの嘔気嘔吐症状の有無を想定しているが、実際には検温表やテンプレート等、今回抽出したテキスト以外に記載された症状は見逃してしまうこととなる。そのため、真の意味で症状の有無を判定しようとする場合は、カルテの記載に限らず症状について記載のあるデータをより広範に分析に加える必要がある。

表 3　ロジスティック回帰の係数

上位10		下位10	
軽度	3.53	ない	-2.75
そう	2.67	気分	-1.99
嘔気 ある	2.60	なし eos	-1.88
嘔吐 eos	2.60	昨日	-1.70
吐き気	2.57	嘔気 嘔吐	-1.53
ある	2.35	なし	-1.48
吐く	2.12	ある て	-1.40
吐き気 ある	2.02	eos bos 吐き気	-1.40
悪心 ある	2.02	気持ち	-1.40
少量	1.87	悪心 嘔吐	-1.40

係数が正の場合は「症状あり」に、負の場合は「症状なし」の方向に判定が傾く。
bos は文頭（beginning of sentence）、eos は文末（end of sentence）に付与されるタグを表す。word 5-gram を用いたため、複数の単語の並びが1語として扱われている。

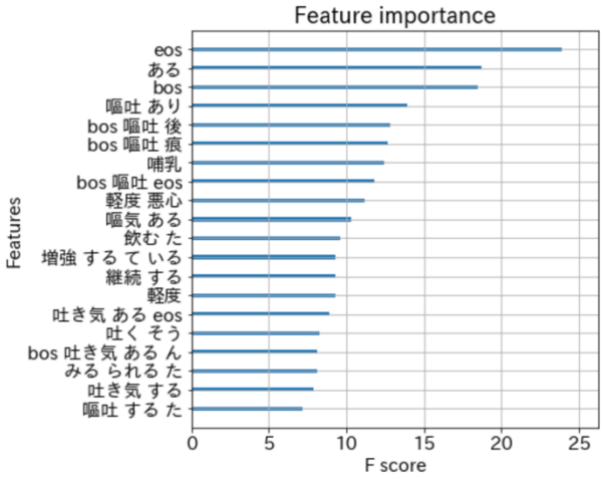

図 3 勾配ブースティング木の変数重要度
上位 20 位までを示す。グラフは XGBoost の plot_importance メソッドを用いて作成し、指標には gain（決定木の分岐においてその特徴量が評価関数の値を改善した大きさに基づく）を用いた。

表 4　嘔気嘔吐のない日を誤って症状ありと判定した例

<BOS> 昨日 夜 カップ ヌードル 食べる た 後 に ちょっと 吐く た <EOS> <BOS> 吐く た すっきり する た <EOS> <BOS> 昨日 夜食 事後 に 少量 嘔吐 あり 胃 部 不快 感 軽減 する て いる <EOS> <BOS> 本日 悪心 嘔吐 ない <EOS> <BOS> 食事 後 に 胃 部 不快 感 嘔吐 認める て いる 為 注意 する て 観察 必要 <EOS>
<BOS> 1 kg / 週 体重 増加 ある 皮下脂肪 増加 考える られる 食欲 減退 軽度 嘔気 みる られる 時 ある ため 腹水 貯留 に 注意 する て 観察 する 必要 ある <EOS> <BOS> 気分 不快 嘔気 出現 ない <EOS>
<BOS> 時々 胃痛 嘔吐 ある <EOS> <BOS> 嘔吐 (-) <EOS> <BOS> 日 によって 嘔吐 胸痛 といった 症状 認める <EOS>

勾配ブースティング木による誤判定。1 例目は過去（昨日）の症状であることを認識できていない。2 例目は「嘔気がみられる時があるため」と言う仮定の表現による誤判定と思われる。3 例目は広い範囲の時期における症状について記載されたもの。

表 5　嘔気嘔吐のある日を誤って症状なしと判定した例

<BOS> 途中 気分 不良 嘔気 訴え あり <EOS>
<BOS> 今 ところ 吐き気 ない <EOS> <BOS> 悪心 咳嗽 時 あり <EOS> <BOS> 急性 副作用 悪心 嘔吐 バイタル 変化 ない <EOS> <BOS> 倦怠 感 継続 する て いる 悪心 出現 ない <EOS>
<BOS> 吐き気 まだ まし <EOS>

勾配ブースティング木による誤判定。1 例目は「嘔気の訴えあり」という表現を学習しきれていない。2 例目は「悪心」と「あり」の間に「咳嗽時」という語句が入ったことにより「悪心」と「あり」の結びつきを検出できなかったものと思われる。3 例目は嘔気の存在が間接的に表現されている。

6. 結論

嘔気・嘔吐をターゲットにした1日単位での症状の有無判定というタスクにおいて、機械学習の手法により複数のアルゴリズムで F1-score 0.80、Precision-Recall AUC 0.89 程度の精度で判定が可能であった。Word N-gram の bag of words によってある程度は否定語の認識も可能であったが、頻度の少ない表現、過去に起きた症状、仮定の表現などについての誤判定が多く、制度の向上にはこれらの解決が課題となる。

参考文献

1) 独立行政法人 医薬品医療機器総合機構. 医療情報データベース「MID-NET」について. 医薬品・医療機器等安全性情報 No.351. 2018 年 3 月.
2) 万病辞書 [http://sociocom.jp/~sociocom/mednlp/index.php/manbyou/].
3) 患者症状抽出器 AEX [http://sociocom.jp/~sociocom/mednlp/index.php/aex/].
4) 工藤拓, 山本薫, 松本裕治. Conditional Random Fields を用いた日本語形態素解析. 情報処理学会研究報告自然言語処理(NL). 2004;2004(47):89-96.

表 6　モデル評価結果

Test Data: N = 2951 (True Positive = 179, True Negative = 2772)

アルゴリズム	Prediction Positive	Prediction Negative	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
ナイーブベイズ	170 (119)	2781 (2721)	0.96	0.70	0.66	0.68	0.98	0.76
ロジスティック回帰	175 (144)	2776 (2741)	0.98	0.84	0.80	0.82	0.98	0.89
勾配ブースティング木	150 (130)	2801 (2752)	0.99	0.86	0.71	0.78	0.99	0.89
サポートベクトルマシン	159 (136)	2792 (2749)	0.98	0.86	0.77	0.81	0.98	0.89
NN（全結合層のみ）	190 (133)	2761 (2715)	0.97	0.70	0.74	0.72	0.98	0.83
NN（1 次元 CNN）	217 (153)	2734 (2708)	0.97	0.71	0.85	0.77	0.98	0.84

NN: ニューラルネットワーク

Prediction Positive と Prediction Negative の( )内は予測が正しかった場合の数

Precision, Recall, F1-score は「症状あり」に対してのもの

テストデータ 2951 のうち 1652 は嘔気嘔吐関連文字列を含まなかったため無条件で「症状なし」と判定されている