一般口演 | 医療データ解析

一般口演17
データマイニング

2021年11月20日(土) 16:30 ～ 18:00 F会場 (2号館2階224)

[3-F-3-02] 電子カルテテキストから構築した症例マトリクスによる協調フィルタリングベースの類似症例予測

*野口 怜[1]、鳥飼 幸太[1]、齋藤 勇一郎[1] （1. 群馬大学医学部附属病院 システム統合センター）
*Rei Noguchi[1], Kota Torikai[1], Yuichiro Saito[1] （1. 群馬大学医学部附属病院 システム統合センター）

キーワード：Discharge summary, Unstructured data, Case matrix, Similar Case Matching, EMR

【背景・目的】電子カルテのテキストデータは、患者の病歴や経過、アウトカムまでの情報が一貫して詳述されており、診療上極めて重要なデータリソースである。しかしながら、非構造化データのため取り扱いが難しくまだ十分に活用されていない。著者らは、電子カルテテキストの診療活用を目指し、退院サマリを用いて、症例ごとに主訴・病歴の出現単語を集計した構造化データ「症例マトリクス」の抽出手法を確立している。本研究では、この症例マトリクスを学習データとして、新たな退院サマリや経過記録テキストを入力として与え、類似症例を検出するフレームワーク構築を目的とした。
【方法】当院における循環器内科などの退院サマリ約10年分（17,000件程度）に対して、匿名化処理後、病名、主訴、既往歴などの情報を正規表現により抽出し、1症例1レコードの形式に構造化した。その後、辞書なしで単語分割可能な自然言語処理エンジン IRIS NLP (InterSystems社) を用いて、症例ごとに単語頻度集計を取ることで症例マトリクスを構築した。なお、表記揺れは、既存の辞書リソースである万病辞書や単語ベクトル化手法を行うことで集約を行なった。
【結果】新たな退院サマリに対して、症例マトリクス構築と同様に、単語頻度集計、表記揺れ集約を行い、症例ベクトル（症例マトリクスの1レコードに相当するデータ）を構築した。この症例ベクトルを入力とし、類似ベクトルを推論する手法「協調フィルタリング」を用いて、症例マトリクスから類似症例を検出するアルゴリズムを構築した。検出された症例には、同一病名の患者も含まれていた。また、同様にして、よりナラティブな文章である経過記録テキストにも適用することができた。
【結語】本研究により、電子カルテテキストを活用した類似症例検索の可能性が示唆された。今後、検査データなどとも組み合わせ、更なる精度向上を図っていきたい。

©一般社団法人 日本医療情報学会

電子カルテテキストから構築した症例マトリクスによる協調フィルタリングベースの類似症例予測

野口 怜[*1]、鳥飼 幸太[*1]、齋藤 勇一郎[*1]
*1 群馬大学医学部附属病院 システム統合センター

Collaborative Filtering-based Similar Case Matching by "Case Matrix" Developed from Text Data in Electronic Medical Records

Rei Noguchi[*1] , Kota Torikai[*1] , Yuichiro Saito[*1]
*1 System Integration Center, Gunma University Hospital

Introduction: Text data recorded in electronic medical records have important and consistent information about patients and are highly expected to be utilized directly for medical examination or diagnosis. We have previously developed a method for extracting structured data "case matrix", in which each record was unique for each case and presence or absence of each symptom was stored in separate columns, from text description in discharge summaries. This methodology enabled us to extract disease-specific feature words for circulatory diseases and to distinguish the diseases from each other. In this study, for further application of case matrix, we develop a framework of similar case matching using Case Matrix based on collaborative filtering algorithm.

Methods: Case matrix were extracted from ten years of discharge summaries in Gunma University Hospital and filtered by the condition that the primary disease name corresponds to eight main circulatory diseases, and finally 1,155 cases was obtained. User-based collaborative filtering was applied to the case matrix, and similar case matching algorithm for an unknown input case was developed.

Results: Similar cases that satisfied a lower limit of correlation coefficient (≥ 0.3) and one of the number of matched words (≥ 3) were able to be extracted for 348 of the 1,155 cases. In addition, in 108 of the 348 cases, the primary disease name of most similar case corresponded to that of the input case. There was no large difference between correlation coefficient and other similarity indexes such as cosine similarity and Hamming distance.

Conclusion: We successfully developed collaborative filtering-based similar case matching algorithm by case matrix. This methodology can be applied to discharge summaries of other hospitals and will enable us to develop cross-institutional huge case matrix and to conduct large-scale similar case matching. We will try to improve the accuracy by removing stop words and incorporating other feature words extracted from other sections in discharge summaries such as "History of Present Illness" and "Family History" into the case matrix.

Keywords: Discharge summary, Unstructured data, Case matrix, Similar Case Matching, EMR

1. 諸言

電子カルテのテキストデータは、患者の病歴や経過、アウトカムに至るまでの情報が一貫して詳述されており、診療上極めて重要なデータリソースである。しかしながら、非構造化データであることや表記揺れの問題などから、取り扱いが難しくまだ十分に活用されていない。著者らは、電子カルテにおけるテキストデータの一次活用を目指し、退院サマリデータを用いて、1 症例ごとに主訴・病歴の出現単語を集計した構造化データ「症例マトリクス」を抽出する手法を確立した [1)]。さらに、この症例マトリクスを活用して、循環器系の疾患判別に繋がる機械学習モデルの構築 [2)] や、合併症パターンの抽出 [3)] に成功している。このように、症例マトリクスには退院サマリの情報が集約化・構造化されているため、疾患判別やパターン抽出を行うための学習データとして大きな活用可能性を秘めている。

一方で、症例マトリクスは、過去当院に入院した症例を網羅的に集積した「症例データベース」としての側面も持ち合わせている。新たな患者の入院サマリや退院サマリから抽出した出現単語列（症例ベクトル）を入力として与え、各症例の出現単語との一致性を評価することで、過去の類似症例を検出できる可能性がある。加えて、一致した出現単語を参照すれば、説明性の高い症例検索が可能となり、症例検索の効率化や精緻化に貢献できる可能性がある。

2. 目的

本研究では、症例マトリクスを活用した類似症例予測を行うためのフレームワーク構築を目的とし、マーケティング分野で頻用される協調フィルタリングアルゴリズムを用いて、新たな出現単語列（症例ベクトル）と類似する症例を症例マトリクスから抽出することを試みる。

3. 方法

3.1 対象データ・疾患

群馬大学医学部附属病院の電子カルテに保管されている、主に循環器内科、呼吸器・アレルギー内科、消化器・肝臓内科などの退院サマリ約 10 年分（2009.7.1~2018.3.31）を抽出した。このうち、主病名が 8 種の心疾患（狭心症 / 虚血性心疾患 / うっ血性心不全 / 急性心筋梗塞 / 慢性心不全 / 持続性心房細動 / 発作性心房細動 / 心房粗動）である症例に絞り込み、さらに複数回入院した患者を除外し、最終的に 1,155 症例を分析対象として得た。

3.2 退院サマリデータの構造化

上記の退院サマリを匿名化した上で Python（version 3.8.6）の分析環境（Jupyter Notebook）に読み込み、正規表現（文字列のパターン抽出を行うための表記法）により、診断病名や主訴、既往歴、現病歴といった項目ごとにテキストを分割して 1 症例 1 レコードの形式に構造化した（図 1）。

なお、病名も自由テキストで記述されているため、正規表現により複数病名を 1 病名ずつに分割後、「万病辞書（ver.202106）」[4] を用いて標準病名に変換し名寄せした。

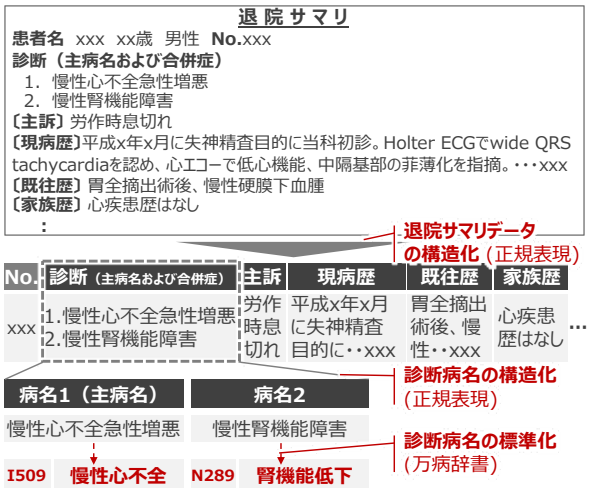

図 1 退院サマリの例とデータ前処理の流れ

3.3 症例マトリクスの構築と類語集約

構造化された退院サマリに含まれる主訴・既往歴・家族歴の自由テキスト記述を対象に、米 InterSystems 社の IRIS NLP[5]を用いて、「エンティティ」と呼ばれる単語単位を抽出した。エンティティは、助詞や動詞などの文法構造をもとに抽出される単語の単位であり、辞書に依存せず、文法レベルでまとまった単語クラスター（主に名詞句）を抽出できることが特徴である。これにより、辞書を整備しなくても専門用語の高精度な抽出が可能となり、また、修飾語が付属したより特異性の高い特徴語が抽出できる。抽出されたエンティティをもとに、症例ごとに出現エンティティの集計を取り「症例マトリクス」を構築した（図2）。

なお、自由テキストには様々な表記揺れ（「咳」「咳嗽」「せき」など）が含まれるため、前述の病名辞書リソースである万病辞書を用いることで代表表現（icd10 における標準表現）に集約を行なった。

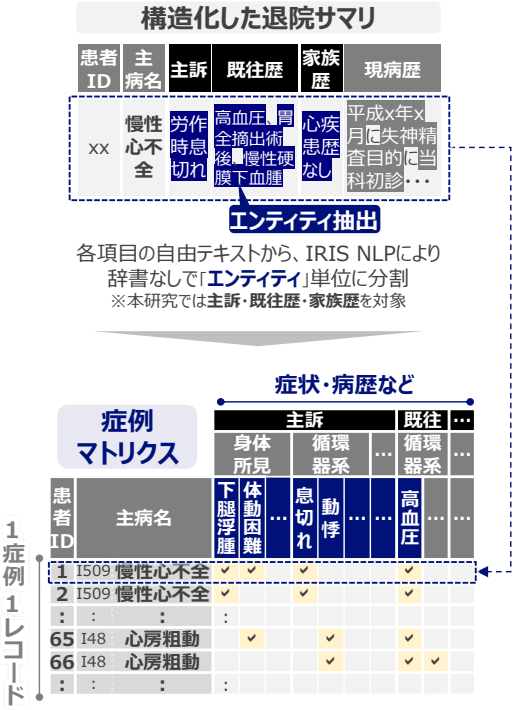

図 2 構造化した退院サマリからの
エンティティ抽出と症例マトリクスの構築

3.4 協調フィルタリング

協調フィルタリングは、主にマーケティング分野において、EC サイトでのパーソナライズしたレコメンデーションエンジンとして古くから活用されている手法である[6)7)]。アイテムに着目して、アイテム間の類似度から類似アイテムを推薦する「アイテムベース」の手法と、ユーザーに着目し、ユーザー間の類似度から、類似ユーザーの嗜好は近いという前提に立ってアイテムを推薦する「ユーザーベース」の手法がある。本研究における類似症例予測への適用を考えると、ユーザーは患者（症例）、アイテムは出現単語（症状）とみなすことができるため、類似症例の抽出が可能な「ユーザーベース」の手法を採用した。

本研究で構築した協調フィルタリングに基づいた類似症例予測の概要を図 3 に示す。予測対象とするある 1 症例から構築した症例ベクトルを入力として、症例マトリクス内の各症例との相関係数を網羅的に算出した。なお、各症例のベクトルは、症状有無を 0, 1 で表現する one-hot ベクトル形式になっているが、0 の値も含めた全体の相関係数として算出した。ただし、一致単語がほとんど存在せずに高い相関を示すような症例を除外するために、「一致単語数」を集計し、これが 3 以上になる症例のみに絞り込んだ。

類似度の指標は、相関係数（ピアソンの積率相関係数）のほか、ベクトル同士の成す角度の近さを表す「コサイン類似度」および、2 進数間の異なる桁数を抽出する「ハミング距離」（例えば、11111 と 10101 とのハミング距離は 2）も用いて、指標の検討も行なった。

類似症例の妥当性評価として、一致した単語の妥当性に加え、予測に用いていない「主病名」の一致性を確認した。

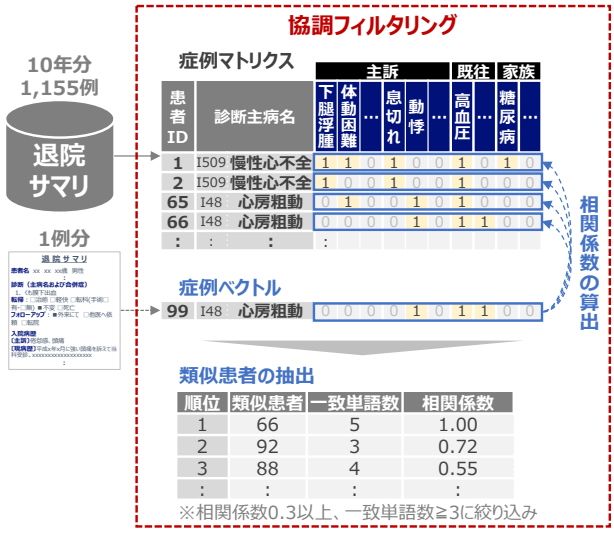

図 3　協調フィルタリングベースの類似症例予測

4. 結果

4.1 類似症例の抽出結果

症例マトリクスに格納された 1,155 症例から 1 症例ずつ取り出し、構築したアルゴリズムを用いて、症例マトリクス内から逐次、類似症例の抽出を試みた。1,155 症例のうち、「相関係数が 0.3 以上かつ、一致した単語数が 3 個以上」という条件を満たす類似症例が見つかったのは 348 症例であった。

また、入力とした症例の主病名と、最も類似度の高い症例の主病名とを突合し、その一致性を確認した。結果として、348 症例中の 108 症例において、類似症例と主病名が一致していることが確認できた。

類似度指標間の違いについては、相関係数とコサイン類似度ではあまり大きな差異は見られなかった。ハミング距離については、相関係数およびコサイン類似度とは異なる結果が得られたが、上位にあがった症例（ハミング距離がより近い症例）の妥当性は低かった。。

4.2 個別の類似症例抽出結果

表 1 に、ある症例ベクトルに対する、類似症例の抽出結果を示す。入力した症例ベクトルの主病名は「慢性心不全」であり、最も類似している症例 ID173 の主病名と一致した。一致した特徴語では、主要な症状や既往歴が得られた。次点の類似症例 ID1704 では、主病名は持続性心房細動であり不一致であったが、一致した特徴語を見ると、既往歴として「慢性心不全」が得られていた。

表 1　ある症例ベクトルに対する類似症例の抽出結果

#	症例 ID	相関係数	特徴語の一致数	一致した特徴語	主病名（標準病名）
0	173	0.327	5	主訴_呼吸困難 既往_高血圧症 既往_糖尿病 既往_心房細動 既往_陳旧性心筋梗塞	慢性心不全
1	1704	0.311	3	既往_高血圧症 既往_糖尿病 既往_慢性心不全	持続性心房細動

5. 考察

5.1 抽出結果の妥当性

1,155 症例のうち、条件を満たす類似症例を抽出できたのは 348 症例（約 30%）であったが、抽出できなかった残りの症例の多くは、とりわけ「1」の値が少ないスパース（疎）な行列であった。このため、相関係数および単語の一致数ともに、低い値が出やすいものと思われ、変数の不足が疑われる。本研究では、症例マトリクスの次元が大きくなりすぎないように、合併症の情報や、長文で書かれやすい現病歴テキストなどは分析対象から除外しているが、これらのデータも変数として症例マトリクスに含めることで、マトリクス中の「1」の値を増やせる可能性がある。一方で、これによりマトリクスのスパース性はより増すことになるため、後述する次元削減手法との組み合わせも検討していく必要がある。

個別の類似症例抽出結果（表 1）では、主訴に「呼吸困難」という慢性心不全における主たる症状の 1 つが含まれていることに加え、既往症にあがった「高血圧症」や「糖尿病」は日本循環器学会のガイドライン [8] でも示されている慢性心不全の高リスク要因であり、妥当な抽出結果であると言える。次点の類似症例においては、主病名こそ合致していなかったものの、既往歴に「慢性心不全」が含まれており、ベクトルの傾向として類似していたことが窺える。

5.2 本研究の意義と今後の発展性

本研究で確立した方法論は、データ加工やモデル適用のアプローチが主眼であり、元のデータ形式や施設に依らず、汎用的に適用できるものである。中でも、学習データやデータベースとして極めて実用性の高い症例マトリクスというコンセプトや、他分野ではレコメンデーション技術として十分に成熟したアルゴリズムである協調フィルタリングを医療データに適用した点は、非常に意義深い。

本方法論を他病院のデータにも適用すれば、多施設の退院サマリデータを「大規模症例マトリクス」として統合することも可能であり、実現すれば、膨大な学習データ、症例データベースとして機能することが期待できる。

また、本研究では、予測対象の入力ベクトルにも退院サマリデータを用いたが、今後は、入院サマリや診察記事から抽出したベクトルを入力として類似症例を参照することで、将来発症しうる症状や合併症の予測にも活用できる可能性がある。

5.3 課題

5.3.1 一致単語数の少なさと妥当性

本研究で抽出された類似症例の中には、相関係数 0.6 以上の比較的高い相関を示す症例も存在した。しかしながら、一致した出現単語は 3~5 個程度にとどまることがほとんどであり、類似性を評価するには物足りなさもある。今回、可能な限り症例マトリクスの次元数を抑えるために、現病歴の出現単語や合併症の情報は除外しているが、これらを改めて変数として活用することで、一致単語数が増加する可能性がある。

また、家族歴に関する一致単語では、「家族_母」「家族_父」などの疾患とは本質的に関係のない単語が混じっており、予測に悪影響を与えている可能性が高い。今後、こうしたノイズとなる単語（ストップワード）は、定義・除外していくことが必要である。

5.3.2 コールドスタート問題

一般に、協調フィルタリングの手法では、情報蓄積の少ない事例について類似検索が難しくなる「コールドスタート問題」[9] を抱えている。今回のケースにおいても、予測対象のデータに新しい出現単語がある場合や、症例の少ない疾患では、類似症例の精度が大きく下がることが想定される。

対応策として、データ数(症例数)の増加、項目値(出現単語)のカテゴリ化などの古典的なアプローチに加え、クラスタリング手法と組み合わせた手法[10] や半教師あり学習[11] が提案されている。

5.3.3 スパース性

本研究で用いた症例マトリクスや、予測対象として入力される症例ベクトルは、膨大なカラムを持つ高次元データであり、さらにその要素の大部分が「0」というスパース(疎)な特徴を持つ。類似度算出に用いた相関係数やコサイン類似度などの指標は、これら膨大な 0 羅列の類似性を含んだものとなってしまっており、精度低下につながっている可能性がある。

スパース性への一般的な対応策は、あらかじめ機械学習手法などを用いてデータの特徴をモデル化する「モデルベース手法」が主流となっているが、中でも、より低次元の二つの非負値行列の積で近似して次元削減する「非負値行列因子分解(Non-negative Matrix Factorization; NMF)」[12] の手法が近年よく協調フィルタリングに適用されている[13]。

6. 結言

退院サマリテキストから構築した症例マトリクスを活用し、協調フィルタリングのアプローチを応用することで、出現単語に基づいた類似症例予測の手法が構築できた。今後は、モデルベース手法との組み合わせも検討し、更なる精度向上を図っていくとともに、症状や合併症の予測といった機能向上に取り組んでいきたい。

参考文献

1) 野口怜, 鳥飼幸太, 齋藤勇一郎. 電子カルテデータを構造化した症例マトリクスによる疾患判別モデルの構築と表記揺れ集約の効果. 第 24 回日本医療情報学会春季学術大会, 2020 年 6 月.
2) 野口怜. 電子カルテデータ構造化による症例マトリクスと疾患判別モデルの臨床導入意義. 月刊新医療 2021 ; 48 : 45-49
3) 野口怜, 鳥飼幸太, 齋藤勇一郎. 退院サマリデータ構造化に基づく合併症パターンの分析. 医療情報学 2020 ; 39 (Suppl.) :663-66.
4) Ito K, Nagai H, Okahisa T, et al. J-Medic: A Japanese disease name dictionary based on real clinical usage. 11th International Conference on Language Resources and Evaluation. 2018 ; 2365-69.
5) Bronselaer G, De Tre. Concept-relational text clustering. journal of intelligent systems. 2012 ; 27 : 970-93.
6) Sarwar B, Karypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. Proceedings of the 10th international conference on World Wide Web. 2001 ; 285-295.
7) Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 2003 ; 7(1) : 76-80.
8) 日本循環器学会/日本心不全学会 合同研究班. 急性・慢性心不全診療ガイドライン. 日本循環器学会/日本心不全学会, 2017.
9) Ricci F, Rokach K, Shapira B . Recommender Systems Handbook Second Edition. Springer 2015.
10) 大西佑紀, 菅谷信介. コールドスタート問題解決のためのソフトクラスタリング活用の提案. 2017 ; (1) : 229-230.
11) Zhang M, Tang J, Zhang X, Xue X. Addressing cold start in recommender systems: A semi-supervised co-training algorithm. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (SIGIR '14). 2014 : 73-82.
12) Lee D, Seung H. Learning the parts of objects by non-negative matrix factorization. Nature. 1999 ; 401 : 788-91.
13) Xin L, Mengchu Z, Yunni X, Qingsheng Z. An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems. IEEE Transactions on Industrial Informatics. 2014 ; 10(2) : 1273-84.