6N-08

頻繁医療指示パターンを用いた医療機関間のクラスタリング手法の提案

澤村 今日子 † 杉谷 美和 † 松尾 亮輔 † 山崎 友義 † 荒木 賢二 †
小口 正人 † 横田 治夫 ‡ Le Hieu Hanh†
お茶の水女子大学 † 城西大学 ‡†

1 はじめに

近年，特定疾病に対する医療指示の比較のために，複数医療機関のクラスタリング分析が行われている．安光らの研究では，複数医療機関の電子カルテデータにシーケンシャルパターンマイニング (SPM) を適用し，それぞれの医療機関の頻出医療指示パターンを抽出し，それらに対してクラスタリングを行い，同じグループに所属されたシーケンスに対して生成された併合シーケンスバリアントのテキストファイルが算出された [1]．しかし，この研究では一つの医療機関に付き一つの代表的な頻出医療指示で医療機関間の距離を計算しているため，医療機関の特徴を充分に表現できなく，クラスタリングの精度が不十分であった. そこで，本研究は医療機関に抽出された全ての頻出医療指示パターンを対象に，医療機関間の距離を計算したより正確なクラスタリング手法を提案する.

2 提案手法

2.1 頻出医療指示パターンの算出方法

これは SPM の手法を拡張したものであり，アイテム間の時間間隔が重要視され，時間依存性のある頻出医療指示パターン（以降頻出シーケンス）を抽出する点に特徴がある．ここで杉谷らの研究 [2] で提案された手法を用いる．

定義 1. 医療シーケンスデータベースからある医療機関 H に対して, 時間間隔を考慮した頻出シーケンス集合 $fs(H)$ に入る頻出シーケンス $fs(h)$ は以下の形式で表される．$fs(h) = \langle(a_1, x_1), (a_2, x_2), \ldots, (a_n, x_n)\rangle$ ここで，$x_j = t_{j+1} - t_j$ は医療指示 a_j と a_{j+1} の間隔を表す.

2.2 再現率・適合率・F 値

第 2.1 で抽出された頻出シーケンスを医療現場に作成された標準クリニカルパスと比較し再現率・適合率・F 値を算出．F 値が最も高くなった minsup を該当医療機関に対する適切な minsup とする.

2.3 動的時間伸縮法による距離計算

医療機関 H と医療機関 K の頻出シーケンスの集合をそれぞれ $fs(H)$ と $fs(K)$ とする．ここで，各集合の要素は定義 1 で定義されたものである．ここで，医療機関 H との距離は動的時間新宿法を用いて以下のように計算される.

$$distance(H,K) = \frac{\sum d(fs(h_p), fs(k_q))}{\parallel fs(H) \parallel \times \parallel fs(K) \parallel}$$

ここで，$fs(h_p) \in fs(H)$, $fs(k_q) \in fs(K)$, $\parallel fs(H) \parallel$ と $\parallel fs(K) \parallel$ はそれぞれ $fs(H)$ と $fs(K)$ の頻出シーケンス数である．そして，$d(fs(h_p), fs(k_q))$ は $fs(h_p)$ と $fs(h_q)$ の動的時間伸縮（DTW）距離である.

$fs(h_p)$ と $fs(k_q)$ を以下のように表し，各シーケンス内の要素（医療指示と間隔の対）がマッ

Proposal of a clustering method for medical institutions using frequently occurring medical instruction patterns

† Kyoko Sawamura†, Miwa Sugitani†, Ryosuke Matsuo†, Tomoyoshi Yamazaki†, Kenji Araki†, Masato Oguchi†, Haruo Yokota‡, Hieu Hanh Le†
†Ochanomizu University ‡Josai University

チしたら 0 で、マッチしなかったら 1 とし，シーケンス間の DTW 距離 $d(fs(h_p), fs(k_q))$ を求める．

$$fs(h_p) = \langle (a_{h_1}, x_{h_1}), (a_{h_2}, x_{h_2}), \ldots, (a_{h_p}, x_{h_p}) \rangle$$

$$fs(k_q) = \langle (a_{k_1}, x_{k_1}), (a_{k_2}, x_{k_2}), \ldots, (a_{k_q}, x_{k_q}) \rangle$$

2.4 階層型クラスタリング

第 2.3 節で計算した頻出シーケンス間の距離と医療機関間の距離を用いて，既存研究と同様に階層型クラスタリングを行う．ただし，距離を用いた他クラスタリング手法の適用も可能である．

3 実験

本研究では，提案手法により抽出された全ての頻出シーケンスを用いたクラスタリング結果と，既存手法である，各医療機関で抽出された頻出シーケンスをマージして得られた単一のシーケンスを用いたクラスタリング結果とを比較する．マージ手法の一例として，[day1: 検査] と [day0: 投薬，day2: 点滴] という 2 つのシーケンスが存在する場合，時間的順序を考慮した上で，[day0: 投薬，day1: 検査，day2: 点滴] のように統合する．クラスタリング結果の精度評価にはコーフェン相関係数を使用した．クラスタリングは，代表的な手法として，ワード法，単純連結法，完全連結法，平均法，重心法の 5 つを比較した．

3.1 データセット

実データを用いた検証では，27 の医療機関から 2015 年から 2024 年までに収集された匿名加工済み電子カルテデータセットを使用した．本研究は一般社団法人ライフデータイニシアティブの利用目的等審査委員会による審査を受け，承認された上で研究を進めた（審査番号 No.2024_MIL_0004_A001）．

3.2 実験結果

結果は図 1 に示した通りである．このように，全ての手法において本提案手法の方がコーフェン相関係数が平均 0.12 高く，より精度の高いクラスタリングを行えることがわかったため，本提案手法の有効性が確認できた．

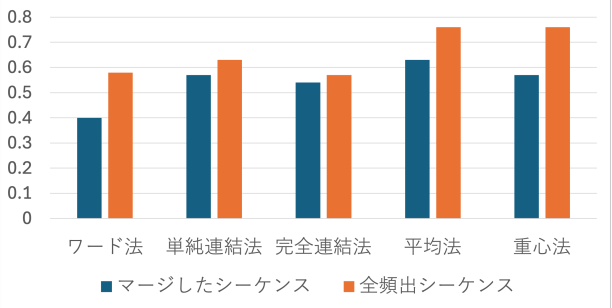

図 1　コーフェン相関係数の結果

4 おわりに

提案手法では，各医療機関から抽出された全ての頻出シーケンスを対象とし，シーケンス間の距離を DTW 法により算出し，それらを平均することで医療機関間の距離を定義した．実データを用いた評価実験では，全頻出シーケンスを用いたクラスタリング結果の方が，頻出シーケンスをマージして得られた併合シーケンスバリアントを用いた場合と比較して，コーフェン相関係数が高く，より高精度なクラスタリングが実現できることを確認した．今後の課題として，データセットに存在する他疾患に対しても本手法を適用し，有効性を確認する．

謝辞

本研究の一部は日本学術振興会科学研究費 (#24K02943) の支援によって行われた．

参考文献

[1] 安光ら．クラスタリングを用いた多病院間の頻出医療指示パターン比較．In *DEIM Forum*, No. 5b-6-3, 2023.

[2] 杉谷ら．複数医療機関間の電子カルテデータを用いた統計情報付き頻出医療指示パターンの抽出と可視化．In *DEIM Forum*, No. 6K-02, 2025.