動的時間伸縮法に基づく頻出医療指示パターン間距離を用いた複数医療機関クラスタリング手法の評価

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[††††]

† お茶の水女子大学情報科学コース　〒 112–8610 東京都文京区大塚 2–1–1
†† お茶の水女子大学理学部　〒 112–8610 東京都文京区大塚 2–1–1
††† 城西大学理学部　〒 102–0093 東京都千代田区平河町 2–3–20
†††† お茶の水女子大学共創工学部　〒 112–8610 東京都文京区大塚 2–1–1
E-mail: †g2120518@is.ocha.ac.jp

あらまし　近年，医療行為の支援のために，単一医療機関に蓄積された電子カルテデータの分析が活発に行われている．電子カルテデータを分析することで，医療処置の流れの差異の評価，医療処置の差異要因の推定，特定の疾患患者に対して行われる典型的なオーダの時系列の作成などが可能となる．さらに複数医療機関の電子カルテデータが分析できるようになって，医療機関から抽出された頻出医療指示パターンを比較することで，特定疾病に対する医療行為の比較をより理解しやすくすることができるが，対象とした複数医療機関が増えると共通パターンが消えてしまうため，効率的に複数医療機関をクラスタリングして，類似した医療機関間で比較する必要がある．しかし，既存手法では，各医療機関につき一つの代表的な頻出医療指示パターンのみを用いて医療機関間の距離を計算しているため，医療機関が有する多様な医療指示の特徴を十分に反映できないという課題がある．また，クラスタリング結果が評価されていないという課題もある．そこで本研究では，各医療機関において抽出された全ての頻出医療指示パターンを対象とし，医療機関間の距離をより適切に算出することで，医療機関の特徴をより正確に反映したクラスタリング手法を提案する．具体的には，医療機関の電子カルテデータに対してシーケンシャルパターンマイニングを適用し，各医療機関の頻出医療指示パターンを抽出する．次に，医療指示手順の時系列構造を考慮するため，動的時間伸縮法に基づいて頻出医療指示パターン間および医療機関間の距離を計算する．最後に，算出した医療機関間距離を用いて，階層型クラスタリング手法と組み合わせた医療機関のクラスタリングを行い，その結果を評価する．評価実験では，実際の医療指示データを用いて，提案手法による距離計算の有効性およびクラスタリング結果の安定性を検証する．

キーワード　クラスタリング，クリニカルパス

1　はじめに

近年，医療行為の支援のために，単一医療機関に蓄積された電子カルテデータの分析が活発に行われている．電子カルテデータを分析することで，特定の疾患患者に対して行われる典型的な頻出医療指示の時系列の作成 [1, 2, 3]，医療処置の流れの共通・差異点の評価 [4, 5]，医療処置の差異要因の推定 [6] などが可能となる．

さらに複数医療機関の電子カルテデータが分析できるようになって，医療機関から抽出された頻出医療指示パターンを比較することで，特定疾病に対する医療行為の比較をより理解しやすくすることができる．しかし，対象とした複数医療機関が増えると共通パターンが消えてしまうため，効率的に複数医療機関をクラスタリングして，類似した医療機関間で比較する必要がある．

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を適用し，各医療機関における頻出医療指示である，シーケンスバリアント（以下 SV）を抽出した．さらに，これらの SV に対して併合シーケンスバリアント（以下 MSV）を生成し，得られた MSV に対してクラスタリングを行い，複数医療機関の頻出医療指示パターンを比較した [7].

しかし，この手法では，各医療機関につき一つの代表的な頻出医療指示パターンのみを用いて医療機関間の距離を計算しているため，医療機関が有する多様な医療指示の特徴を十分に反映できないという課題がある．その結果，医療機関間の類似性を正確に捉えることが難しく，精度の良いクラスタリングが実現できない可能性は高い．また，クラスタリング結果が評価されていないという課題もある．

そこで本研究では，各医療機関において抽出された全ての頻出医療指示パターンを対象とし，医療機関間の距離をより適切に算出することで，医療機関の特徴をより正確に反映したクラスタリング手法を提案する．具体的には，医療機関の電子カルテデータに対してシーケンシャルパターンマイニングを適用し，各医療機関の頻出医療指示パターンを抽出する．次に，医療指

示手順の時系列構造を考慮するため，動的時間伸縮法（DTW）に基づいて頻出医療指示パターン間および医療機関間の距離を計算する．最後に，算出した医療機関間距離を用いて，階層型クラスタリング手法と組み合わせた医療機関のクラスタリングを行い，その結果を評価する．評価実験では，実際の医療指示データを用いて，提案手法による距離計算の有効性およびクラスタリング結果の安定性を検証する．

以下に本研究の貢献を挙げる．

- 複数医療機関のクラスタリングのために，DTW 法を用いた医療機関間の距離の計算方法を提案する．
- 27 医療機関の実データを利用した，シルエットスコアと Dunn 指標を用いた階層クラスタリングの精度評価を通じて，既存クラスタリング手法より安定性向上の確認ができた．

本論文は以下の通り構成される．第 2 節では背景知識について述べ，第 3 節では提案手法であるクラスタリングアルゴリズム，評価方法について説明する．第 4 節では実際に提案手法を用いて実際の複数医療機関のクラスタリング結果を報告し，最後に第 5 節でまとめと今後の課題について述べる．

2 背景知識と関連研究

2.1 クリニカルパス

日本クリニカルパス学会の定義によると [8]，患者状態と診療行為の目標，および評価・記録を含む標準診療計画であり，標準からの偏位を分析することで医療の質を改善する手法のことである．また，それとは別に電子クリニカルパスというのもある．電子クリニカルパスとは情報通信技術を用いて標準診療計画を作成し，標準診療計画に基づく診療の実施を支援し，患者個別の診療状況とその評価を記録し，逸脱事例の集計と分析などを処理する医療管理手法のことである．

2.2 頻出医療指示の抽出方法

T-PrefixSpan[2] では，電子カルテデータから疾患ごとの医療指示シーケンスを抽出し，日付情報を基に時間間隔を考慮した頻出医療指示シーケンスとして頻出クローズドパターンを抽出する．本論文では，杉谷らの研究によって提案された，T-PrefixSpan を用いた疾患ごとに頻出する医療指示パターンを抽出する手法を用いる [9]．これには，以下に示す定義を使用する．

定義 1 (同日の医療指示の並び替えルール)．医療指示の実行時刻がないという前提で，同一日に複数の医療が記録されている場合は，以下の順序で並び替えを行う．

$$\begin{bmatrix} 1. & \text{手術} \\ 2. & \text{投薬} \\ 3. & \text{検査} \\ 4. & \text{診療行為} \end{bmatrix}$$

さらに，同一カテゴリ内で複数の医療指示が存在する場合は，事前に作成した辞書に基づき辞書順（あいうえお順）で並び替える．このルールにより，医療指示の順番が統一され，頻出医療指示パターン抽出の精度向上が期待される．

定義 2 (医療指示シーケンス)．患者 p_i に対する医療指示のシーケンスを以下のように定義する．

$$S_{p_i} = \langle s_1, s_2, \ldots, s_n \rangle$$

s_j は j 番目の医療指示を表し， $s_j = (t_j, a_j)$ である．なお，t_j は医療指示が実施された経過日数を表し，a_j は医療指示の種類（例：手術，投薬，検査など）を表す．

定義 3 (医療シーケンスデータベース， MSDB)．医療シーケンスデータベース (MSDB) は複数患者の医療指示列から構成される．MSDB 内のデータ集合 D は以下のように定義される．

$$D = \{(SID_1, S_{p_1}), (SID_2, S_{p_2}), \ldots, (SID_m, S_{p_m})\}$$

ここで，SID_i は i 番目の患者の識別子であり， S_{p_i} はその患者における医療指示シーケンスを示す．

定義 4 (T-PrefixSpan による頻出シーケンス)．MSDB 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．

2.3 動的時間伸縮法（DTW 法）

櫻井らの論文によると，[10]2 つのシーケンス間の DTW 距離とは，それらのシーケンス長の調整後の距離の和である．長さ n のシーケンス $A = (a_1, a_2, ..., a_n)$ と長さ m のシーケンス $B = (b_1, b_2, ..., b_m)$ を考えた時に，これらの DTW 距離 $DTW(A, B)$ は以下のように定義される．

$$f(t,i) = ||a_t - b_i|| + \min \begin{cases} f(t, i-1) \\ f(t-1, i) \\ f(t-1, i-1) \end{cases}$$

ここで，$f(0,0) = 0$, $f(t,0) = f(0,i) = \infty$, $(t \in \{1..n\}, i \in \{1..m\})$ である．

ここで $||a_t\text{-}\, b_i||$ は 2 つの数値の距離を表す．

なお，本研究ではシーケンスを比較し同じ文字列であれば距離を 0，違う文字列であれば 1 とする処理を加えた後に上記の計算を行う．

$\langle (0, \text{手術}), (1, \text{投薬}), (2, \text{検査}) \rangle$ と $\langle (0, \text{手術}), (1, \text{検査}), (2, \text{投薬}) \rangle$ の 2 つのシーケンスの DTW 距離を計算するとする例を図 1 に表す．まず 1 つ目の $(0, \text{手術})$ は共通であるから，距離は 0 となる．次に 2 つ目の $(1, \text{投薬})$ と $(1, \text{検査})$ は異なるため，距離は 1 となる．同様に 3 つ目の $(2, \text{検査})$ と $(2, \text{投薬})$ も異なるため，距離は 1 となり最終的な DTW 距離は 2 となる．

B/A	day0:手術	day1:投薬	day2:検査
day0:手術	0	1	2
day1:検査	1	1	2
day2:投薬	2	2	2

図 1　2 つのシーケンス間の DTW 距離計算の例

2.4　クラスタリング評価指標

2.4.1　シルエットスコア

シルエットスコアとは，クラスタリング精度を測る指標である [11].

データ集合 D が $C_1, C_2 ... C_k$ の k 個のクラスタに分けられているとする．ここで，以下の 2 指標を計算する．

$$a(o) = \frac{\sum_{o' \in C_i, o' \neq o} dist(o, o')}{|C_i| - 1}$$

$$b(o) = \min_{C_j : 1 \leqq j \leqq k, j \neq i} \left\{ \frac{\sum_{o' \in C_j} dist(o, o')}{|C_j|} \right\}$$

$a(o)$ は各データ点 o に対し，o が属するクラスタ C_i 内の，他の全ての点との平均距離を表し，$b(o)$ は．オブジェクト o が属していないすべてのクラスタに対する平均距離のうち最小の値を表し，o が他クラスタからどの程度分離されているかを示す．

シルエットスコア $s(o)$ は次のように計算する．

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}'}$$

この値の範囲は (-1, 1) である．$s(o)$ の値が 1 に近づくほど o が属するクラスタは，コンパクトかつ他のクラスタから十分に離れていることを意味し，望ましいクラスタリング結果である．一方，値が負になる場合，o は同じクラスタ内のオブジェクトよりも，他クラスタのオブジェクトに近いことを示しており，望ましくないクラスタリング結果である．

2.4.2　Dunn 指標

Dunn 指標とは，クラスタリング精度を測る指標である [12].

データ集合 D が $C_1, C_2 ... C_k$ の k 個のクラスタに分けられているとする．o を各データ点とする．

クラスタの凝縮度 Δ は次のように計算される．

$$\Delta = \max_{C(o_i) = C(o_j)} \{d(o_i, o_j)\}$$

これは，同一クラスタに属する 2 点間の最大距離として定義される．

異なるクラスタの間の分離度 δ は次のように計算される．

$$\delta = \min_{C(o_i) \neq C(o_j)} \{d(o_i, o_j)\}$$

これは，異なるクラスタに属する 2 点間の最小距離として定義される．

Dunn 指標はこれらの比として定義され，次のように計算される．

$$DI = \frac{\delta}{\Delta}$$

この比が大きいほど，クラスタの凝集度に対してクラスタ間の分離度が大きく，より望ましいクラスタリング結果であるといえる．Dunn 指標はクラスタ内およびクラスタ間の距離として最大値・最小値を用いるため，外れ値の影響を受けやすいという特徴がある．

2.5　関 連 研 究

牧原らの研究では，電子カルテの操作ログからクリニカルパスの作成の補助を行い [1]，佐々木らの研究では，クリニカルパス作成のための頻出医療指示の抽出方法を提案した [2]．そして Uragaki らの研究では，作成したクリニカルパスを評価し，新たな分岐や変種を推薦する手法を提案した [3]．Le らの研究では，クリニカルパスの相違点の検出方法を提案し [4]，山田らの研究では，そのような相違点の妥当性および安全性の評価手法を提案した [5]．そして Le らの研究では，それらの相違点の要因分析を行なった [6].

安光らの研究では，3 以上の複数医療機関の頻出医療指示パターンの距離を定義してクラスタリングを行い，対象となる病院群をいくつかのクラスタに分類し，そのクラスタ内で分類を行った [7, 13]．そこで最終結果の併合シーケンスバリアント (MSV) をグラフとして可視化したのだが，MSV だけで医療機関の特徴を表しているため，十分に表すことができていなかった．

また，Valerie らの研究では [14]，順序付きのアイテム集合からのクラスタリングを行ったのだが，一つの要素に対し一つのアイテム集合しかなかったため，今回のように一つの要素に対し複数のアイテム集合が含まれるデータには適用できない．

3　提 案 手 法

本節は提案手法について説明する．まず，杉谷らの手法 [9] を用いた頻出医療指示の算出方法について説明したあとに，その際に使う適切な minsup の算出方法について説明する．次に，DTW を使用した複数医療機関の距離計算の手法を述べる．最後に，クラスタリング手法と評価方法について述べる．

3.1　頻出 SV の算出方法

疾患ごとに頻出する医療指示パターンを抽出する．本研究では，第 2.2 節で説明した，杉谷ら [9] が提案した，疾患ごとに頻出する医療指示パターンを抽出する手法を使用する．ここで，最小支持度（minsup）ごとに抽出されるパターン群が違うため，クラスタリングを対象となるパターンを決める必要がある．そこで，複数の minsup を試して実際の治療に最も近いパターン群を抽出した適切な minsup を決定する．

3.2　適切な minsup の定め方

本研究では，抽出されたパターン群に含まれるパターンを，必要最低限の診療計画であるコアなクリニカルパスと比較し，再現率・適合率・F 値を算出する．そしてその中でも平均 F 値が最も高かった minsup を採用する．

表 1　実 験 環 境

項目	内容
プログラミング言語	Python
ライブラリ	PrefixSpan 実装
データベース	PostgreSQL ver. 16.6
データベースドライバー	Psycopg

適切な minsup を定めるために，抽出された頻出 SV を医療現場に作成された標準クリニカルパスを比較し再現率・適合率・F 値を算出する．

再現率は実際に正の中で正と判断された割合のことで，以下のように計算する．

$$\frac{\text{真陽性}}{\text{真陽性} + \text{偽陰性}}$$

適合率は正と予測した中で実際に正だった割合のことで，以下のように計算する．

$$\frac{\text{真陽性}}{\text{真陽性} + \text{偽陽性}}$$

F 値は適合率と平均率の調和平均のことで，以下のように計算する．

$$\frac{2 \times \text{再現率} \times \text{適合率}}{\text{適合率} + \text{再現率}}$$

以下は再現率・適合率・F 値を計算する例を挙げる．例えば，標準クリニカルパスを [day0: 手術，点滴，day1: 点滴，検査] とする．頻出 SV の [day0: 手術，day2: 検査] に対して，正の医療指示をカバーできたのは全 4 項目のうち [day0: 手術] のみであるので再現率は $\frac{1}{4} = 0.25$ となる．またこの頻出 SV には 2 つの医療指示があるが，正を満たしている医療指示は 1 つなので，適合率は $\frac{1}{2} = 0.5$ となる．よって F 値は $\frac{2 \times 0.5 \times 0.25}{0.5 + 0.25} = 0.33$ となる．

3.3　動的時間伸縮法による距離計算

時系列を考慮できるかつ，医療指示列の要素数が異なっても計算できるため動的時間伸縮法（DTW）を採用した．

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 4 で定義されたものである．ここで，医療機関 H と K の距離は DTW 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 $fs(k_q)$ の DTW 距離である．

$fs(h_p)$ と $fs(k_q)$ を以下のように表すことができるため，各シーケンス内の要素（医療指示と間隔の対）がマッチしたら 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．

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

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

3.4　階層型クラスタリング

第 3.3 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．

3.5　クラスタリング評価方法

医療現場などにより作成された，正確なクラスタリング結果を用いた評価は現段階では困難であるため，今回は内部評価指標である，シルエットスコアと Dunn 指標を使用する．これらの指標については第 2.4.1 節と，第 2.4.2 節で詳しく説明している．

4　実　　験

本節は，提案手法の評価実験について説明する．まず，実験方法および実験環境を述べる．最後に，実験結果について述べる．

4.1　実 験 方 法

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シルエットスコアと Dunn 指標を使用した．

頻出 SV の抽出においては最小支持度 (minsup) は 0.3〜0.8 の幅で行い，第 3.2 節で説明した再現率・適合率・F 値を計算し，そこで算出できた，平均 F 値が最も高いとなった最適な minsup を採用した．

クラスタリング結果の精度評価には，シルエットスコアおよび Dunn 指標を用いた．これらはいずれもクラスタリングの妥当性を評価する指標であり，シルエットスコアは値が 1 に近いほどクラスタ内の凝集度およびクラスタ間の分離度が高いことを示す．

本研究では，クラスタサイズが過剰に小さいクラスタリングを発生させないために，最小クラスタサイズを 3 と定義し，それに満たないクラスタは最も近い別のクラスタにマージする手法を採用した．また，シルエットスコアが最大となり，かつ Dunn 指標が無限大とならないクラスタ数を最適なクラスタ数として決定した．つまりクラスタリングの条件は以下となる．

- 最小クラスタサイズが 3
- Dunn 指標が無限大（Infinity）ではない
- これらを満たさない場合は Dunn 指標を 0，シルエットスコアを-1 とする

クラスタリングは，代表的な手法として，ワード法，単純

表 2　実データセットのスキーマ

テーブル名	データスキーマ	医療指示カテゴリとの関連
患者情報	施設 ID, 患者 ID, 入院日, 年齢, 性別, BMI	属性情報
日付情報	施設 ID, 患者 ID, 日付, 入院日, 退院日	シーケンス長
傷病情報	施設 ID, 患者 ID, 入院日, ICD-10 コード	疾患名
手術情報	施設 ID, 患者 ID, 入院日, 手術日, K コード	手術
薬剤情報	施設 ID, 患者 ID, 入院日, 投与日, 薬効分類コード	投薬
検査情報	施設 ID, 患者 ID, 入院日, 測定日, 検査名, 検査値, 単位	検査
診療行為情報	施設 ID, 患者 ID, 入院日, 実施日, レセプト電算コード	診療行為
バイタル情報	施設 ID, 患者 ID, 入院日, 測定日, 項目名, 測定値, 単位	—
疾患情報	施設 ID, 患者 ID, 入院日, DPC コード	属性情報

表 3　実データセット: 疾患ごとの患者数

疾患名	延患者数（名）
乳房の悪性腫瘍	12,841
狭心症・慢性虚血性心疾患	11,044
肺の悪性腫瘍	10,005
膀胱腫瘍	9,720
胃の悪性腫瘍	5,961
急性心筋梗塞	4,929
肝・肝内胆管の悪性腫瘍	2,016
椎間板変性・ヘルニア	1,194
子宮頸部・体部の悪性腫瘍	888
合計	**58,598**

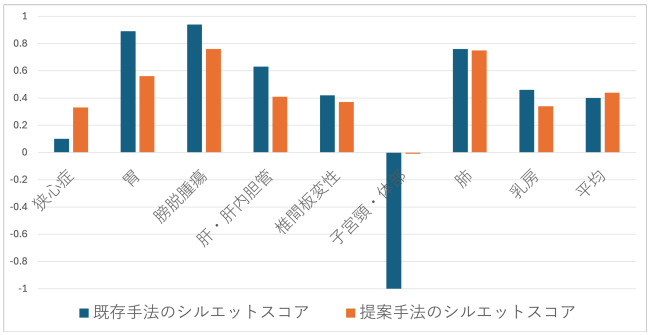

図 2　各疾患のシルエットスコアの結果

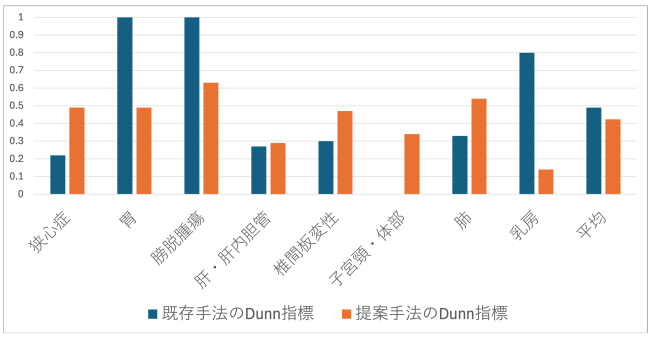

図 3　各疾患の Dunn 指標の結果

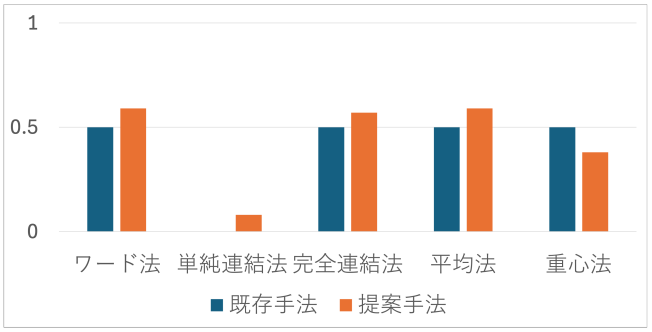

図 4　各クラスタリング手法のシルエットスコアの結果

連結法（Single-link），完全連結法（ Complete-link），平均法（Average-link），重心法（Centroid）の 5 つを比較した．

4.2　実験環境とデータセット

本研究では表 1 に示した通りの環境で実装した．Python のライブラリである「linkage」を使用し，デンドログラムで実装結果を可視化した．シルエットスコアの計算は Python の「sklearn.metrics」ライブラリを使用し，Dunn 指標は「valid-clust」ライブラリを使用した．

実データを用いた検証では，複数の医療機関から 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）．このデータセットは，患者のプライバシー保護のため厳格な匿名化処理が施されており, 異なる医療機関のデータを比較分析できるよう，医療指示コード等の表記統一が行われている．

表 2 に示すのは，本実験で使用した実データセットのスキーマである．ここで，BMI は Body Mass Index の略でボディマス指数である．

表 3 には，実データセットにおける主要な疾患とその患者数を示す．各疾患の患者数には大きなばらつきがあり，これは実際の臨床現場における疾患の頻度やデータ収集状況を反映している．合計で 58,598 名の患者データが含まれている．このデータを用いることで，特定の疾患に偏らない，より現実的な状況下での提案手法の有効性を検証できる．

4.3　実 験 結 果

図 2 と図 3 に示したのは各疾患ごとのシルエットスコアと Dunn 指標の結果である．シルエットスコアの平均は提案手法が 0.439，既存手法が 0.400 となり，提案手法の方が 0.039 高く，Dunn 指標の平均は提案手法が 0.424，既存手法が 0.490 となり，提案手法の方が 0.066 低かった．図 4 と図 5 にに示したのは各クラスタリング手法ごとのシルエットスコアと Dunn 指標の結果である．グラフを見てわかる通り，疾患ごとのシルエットスコア・Dunn 指標に対し，手法ごとの結果では提案手

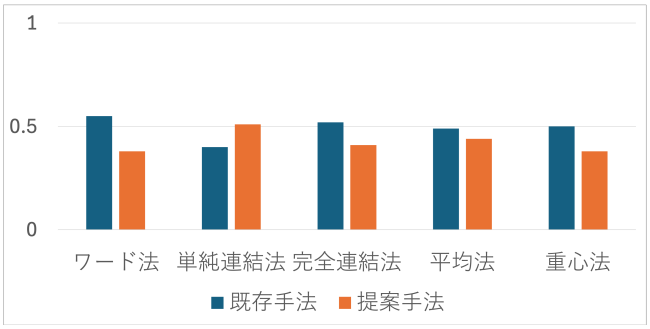

図 5　各クラスタリング手法の Dunn 指標の結果

表 4　クラスタリング条件を満たさなかった事象

疾患名	手法の種類	クラスタリング手法
狭心症，慢性虚血性心疾患	既存手法	単純連結法
子宮頸・体部の悪性腫瘍	提案手法	単純連結法
子宮頸・体部の悪性腫瘍	既存手法	ワード法
子宮頸・体部の悪性腫瘍	既存手法	単純連結法
子宮頸・体部の悪性腫瘍	既存手法	完全連結法
子宮頸・体部の悪性腫瘍	既存手法	平均法
子宮頸・体部の悪性腫瘍	既存手法	重心法
乳房の悪性腫瘍	提案手法	単純連結法
乳房の悪性腫瘍	既存手法	単純連結法

表 5　標 準 偏 差

	提案手法	既存手法
シルエットスコアの標準偏差（疾患別）	0.234	0.588
シルエットスコアの標準偏差（手法別）	0.198	0.200
Dunn 指標の標準偏差（疾患別）	0.147	0.360
Dunn 指標の標準偏差（手法別）	0.048	0.050

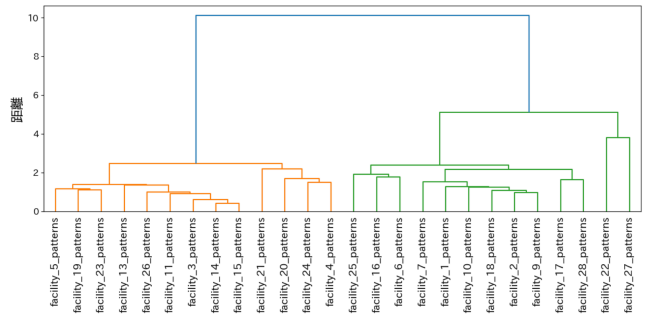

図 6　提案手法によるクラスタリング結果（ワード法）

法と既存手法に大きな差はない．つまり，クラスタリング結果の精度の差は疾患ごとのシーケンスの差によって生まれていることが考察できる．また，手法ごとの結果では提案手法と既存手法共にワード法がシルエットスコア・Dunn 指標共により高くなった．

しかし，既存手法では，クラスタリング条件を満たさない事象が多く発生した．表 4 に示したのがクラスタリング条件を満たさなかった事象である．本事象の 7/9 が既存手法であり，提案手法よりも既存手法の方が，条件を満たさなかった場合が 7.5 倍多いことがわかる．特に，子宮頸・体部の悪性腫瘍は，既存手法ではどの手法でもクラスタリングができなかった．よって，シルエットスコアや Dunn 指標が高く出たのも，一部のクラスタリング結果が上手くいった疾患のみであり，全体的なものではないことがわかる．

また，表 5 で示した通り，全ての結果において提案手法の方が標準偏差が低く，安定したクラスタリング結果であることがわかった．これらの結果から，提案手法の方がより多くの疾患に対し安定的なクラスタリングが行えることがわかった．

実験結果の一例とて，狭心症に関する頻出医療指示に対し，提案手法でワード法を使ってクラスタリングした結果を図 6 に示す．

5　おわりに

5.1　ま と め

本研究では，複数医療機関における頻出医療指示パターンの共通点および相違点をより正確に把握することを目的として，DTW に基づく医療機関間距離を用いたクラスタリング手法を提案した．従来手法では，各医療機関を単一の代表的な頻出医療指示パターンで表現していたため，医療機関の特徴を十分に反映できないという課題があった．また，クラスタリング精度の評価が不十分という課題もあった．

提案手法では，各医療機関から抽出された全ての頻出 SV を対象とし，SV 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．これにより，医療指示の時系列的な違いを考慮した距離計算が可能となった．実データを用いた評価実験では，手法別の結果では提案手法と既存手法どちらもワード法の精度が高かった．また提案手法は，既存手法と比較して，安定したクラスタリング結果を示した．以上より，提案手法は，より汎用性が高く実用的なクラスタリング手法であるといえる．

5.2　今後の課題

今後の課題として，事前に病床数や病院の種類などでクラスタリングを行い，その結果と本手法での結果を比較する．また，本研究を実際に医療従事者の方に発表し，どれだけ現場での実用性があるかを確認する．

謝　　辞

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

文　　献

[1] 牧原健太郎, 荒堀喜貴, 渡辺陽介, 串間宗夫, 荒木賢二, 横田治夫. 電子カルテシステムの操作ログデータの時系列分析による頻出シーケンスの抽出. In *DEIM Forum*, No. F6-2, 2014.

[2] 佐々木夢, 荒堀喜貴, 串間宗夫, 荒木賢二, 横田治夫. 電子

カルテシステムのオーダログデータ解析による医療行為の支援. In *DEIM Forum*, No. G5-1, 2015.

[3] Keishiro Uragaki, Tomoyuki Hosaka, Yoshitaka Arahori, Muneo Kushima, Tomoyoshi Yamazaki, Kenji Araki, and Haruo Yokota. Sequential pattern mining on electronic medical records with handling time intervals and the efficacy of medicines. In *Proceedings of the first IEEE Workshop on ICT Solutions for Health in conjunction with the 21st IEEE International Symposium on Computers and Communications*, pp. 20–25, 2016.

[4] Yuichi Honda, Muneo Kushima, Tomoyoshi Yamazaki, Kenji Araki, and Haruo Yokota. Detection and visualization of variants in typical medical treatment sequences. In *Proceedings of the 3rd International Workshop on Data Management and Analytics for Medicine and Healthcare (DMAH) in conjunction with the 43rd International Conference on Very Large Data Bases*, pp. 88–101, 2017.

[5] 山田達大, 本田祐一, 萱原正彬, Le Hieu Hanh, 串間宗夫, 小川泰右, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. Sid を保持するシーケンシャルパターンマイニングによるクリニカルパスバリアント分析. In *DEIM Forum*, No. D1-1, 2019.

[6] Hieu Hanh Le, Tatsuhiro Yamada, Yuichi Honda, Takatoshi Sakamoto, Ryosuke Matsuo, Tomoyoshi Yamazaki, Kenji Araki, and Haruo Yokota. Methods for analyzing medical-order sequence variants in sequential pattern mining for electronic medical record systems. *ACM Transactions on Computing for Healthcare*, Vol. 4, No. 1, pp. 3:1–3:28, 2023.

[7] 安光夕輝, Le Hieu Hanh, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫. 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.

[8] 日本クリニカルパス学会. `https://www.jscp.gr.jp/index.html`. オンライン, アクセス 2025 年 1 月 8 日.

[9] 杉谷美和, 松尾亮輔, 山崎友義, 荒木賢二, 横田治夫, 小口正人, Le Hieu Hanh. 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.

[10] 櫻井保志, Christos Faloutsos, 山室雅司. ダイナミックタイムワーピング距離に基づくストリーム処理. In *DEWS*, No. L6-5, 2007.

[11] Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. In *Journal of Computational and Applied Mathematics*, pp. 53–65, 1987.

[12] J.C.Dunn. Well-separated clusters and optimal fuzzy partitions. In *Journal of Cybernetics*, pp. 95–104, 1974.

[13] Le Hieu Hanh, Yuki Yasumitsu, Ryosuke Matsuo, Tomoyoshi Yamazaki, and Haruo Yokota. A clustering-based sequence variants analysis method for electronic medical records of multimedical institutions. In *Proceedings of the 7th IEEE International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 653–659, 2024.

[14] Valerie Guralnik and George Karypis. A scalable algorithm for clustering sequential data. In *Proceedings of the first International Conference on Data Mining (ICDM)*, pp. 179–186. IEEE, 2001.