2J3-OS-16b-4

期間付き状態関係に基づく時系列医療データからの頻出パターンマイニング

Frequent Temporal Pattern Mining for Medical Data based on Ranged Relations

平野章二 *1　津本周作 *1
Shoji Hirano　Shusaku Tsumoto

*1 島根大学医学部医療情報学講座
Department of Medical Informatics, Shimane University, School of Medicine

This paper presents a temporal pattern mining method for medical data. It extends Batal's algorithms to handle ranged relations. Experimental results demonstrate that the proposed method could generate frequent patterns with abstracted time ranges embedded in their temporal relations.

1. はじめに

Batal らの提案した時系列頻出パターンマイニング法 [1] は，抽象化と Temporal Logic[2] を採り入れることで，量的なデータと質的なデータが混在する多次元の時系列医療データセットから，イベントの時間関係 (前に生じる／同時に起こるなど) を含む頻出パターンの生成を可能とする有用な手法である。例として，図 1 のように血栓発生をイベント（クラス）とし，血小板数（PLT）と薬 A の投与を属性とするパターンマイニングを考える。Batal らの方法では，まず血小板数と投薬の状態を状

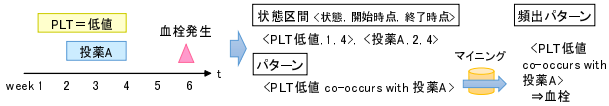

図 1: 状態区間と関係に基づくパターンマイニング.

態区間として記述し，それらを時間的前後関係（before もしくは co-occur with）と組み合わせてパターンを構成する。続いて，同一クラスに属する全データを対象に頻出パターンマイニングのアルゴリズムを適用し，最小サポート基準を満たす頻出パターンを抽出する。それにより,「血小板数が基準下限を下回る状態で薬 A が投与されると，血栓が発生する」というルールを生成することができる。その際に自身のサブパターンよりも有意に予測能力が高い MPTP(Minimal Predictive Temporal Pattern) を選ぶことで，冗長なパターンの生成を抑制できることも大きな特徴である。

しかしながら，この方法で生成されるパターンは，状態の前後関係や共起関係を反映する一方，状態の持続時間や間隔の長短に関する情報を含まない。すなわち，薬 A の投与と PLT 低値の状態がどの程度の期間に渡って継続し，それからどのくらい経過して血栓を生じるかを知ることはできない。このように，イベントに至るまでの，疾患の時間進展を表す知識の獲得が難しいことが課題である。

本稿では，Batal らの方法を拡張し，状態の継続期間や間隔を明示的に組み込んだ時系列頻出パターンマイニング法を提案する。提案法は，before，co-occur with の 2 種類の状態関係を，日，週，年など抽象化された期間を伴う期間付き状態関係へと拡張することで，例えば，'PLT 低値 co-occurs several weeks with 投薬 A'，'PLT 低値 several months before 血栓発生' のような時間進展パターンの記述を試みるものである。

連絡先: 島根大学医学部医療情報学講座 平野章二
〒 693-8501 島根県出雲市塩冶町 89-1
Phone:(0853)20-2173，E-mail:hirano@ieee.org

2. 期間付き状態関係の導入

本節ではまず，基礎となる Batal らの方法について述べる。検査値の高低や投薬の有無などを表す記号 E と，対応する時間区間 (始点 b，終点 e) の情報を組み合わせ，任意の状態を $S = (E, b, e)$ の形式で表現する。例えば図 1 ようにある患者の血小板数が期間 $1 \sim 4$ において低値である状態は，$S = (\mathrm{PLT{=}L}, 1, 4)$ となる。このように表した複数の状態 S_i を開始時刻順に並べることで，任意の時系列を状態系列 $< S_1, S_2, \ldots, S_k >$ として記述できる。パターンはこの状態系列を元に構成されるが，各々の状態が区間を伴うため，その重なりなど関係を定義しなければならない。Batal らは，Allen の論文 [2] で定義される 13 種類の関係の中から，"E_i before E_j"(ある状態 E_i が終わってから別の状態 E_j が始まる：$e_i < b_j$) と "E_i co-occurs with E_j"(ある状態 E_i が始まり，終わる前に別の状態 E_j が始まる：$b_j \leq e_i$) の 2 種類を利用している。この関係を，状態系列に含まれる全ての状態のペア $\{(S_i, S_j) : i < j\}$ に対して割り付けることで，長さ k のパターン (k-pattern という)P_k を，k 個の状態とその関係マトリクス R を用いて $P_k = (< S_1, S_2, \ldots, S_k >, R)$ と表現できる。なお，状態の開始日時と終了日時に関する情報は関係マトリクス生成後には不要となるため，パターン内には保持しない。

頻出パターンの抽出は，候補の生成と，支持度に基づく候補選別の 2 段階で行われる。候補の生成では，長さ k の頻出パターンの先頭に，頻出 1-pattern を 1 つ挿入することで，新たな候補 $(k+1)$-pattern を生成する。例として，図 2 に示すように，V1=L, V2=H, M1=ON という 3 つの頻出 1-pattern があり，'V1=L $\langle b \rangle$ V2=H' という頻出 2-pattern から候補 3-pattern を生成するケースを考える。簡単のため，以降は関係 before と co-occur をそれぞれ $\langle b \rangle$, $\langle c \rangle$ と略記する。まず，この 2-pattern の先頭に，M1=ON を挿入する。新たに挿入した M1=ON と既にある 2 つの状態との関係は未定であるため，可能な関係の組み合わせ $((\langle b \rangle,\langle b \rangle), (\langle c \rangle,\langle b \rangle),(\langle c \rangle,\langle c \rangle)$ の 3 種類) を列挙して関係マトリクスの先頭行を埋めることで，3 つの候補 3-pattern が生成される。なお，関係の組み合わせのうち，$\langle b \rangle$ の後に $\langle c \rangle$ が続くもの $(\langle b \rangle, \langle c \rangle)$ については時間的な論理矛盾となるため候補から除外される。この 3 つの候補パターンの内，支持度が閾値を上回るものが頻出 3-pattern となる。

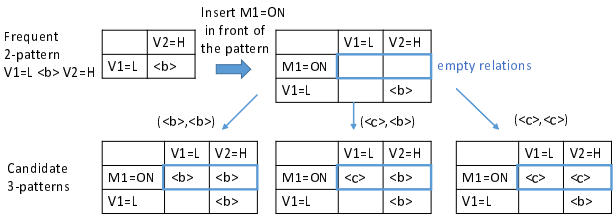

図 2: 候補パターンの生成.

表 1: 期間付き状態関係.

Relation	Meaning	Range d_{ij}[days]
$\langle cd \rangle$	co-occur several days with	$[-7, 0]$
$\langle cw \rangle$	co-occur several weeks with	$[-31, -7)$
$\langle cm \rangle$	co-occur several months with	$(-inf, -31)$
$\langle bd \rangle$	several days before	$(0, +7]$
$\langle bw \rangle$	several weeks before	$(+7, +31]$
$\langle bm \rangle$	several month before	$(+31, +inf)$

提案法では，新たに「期間付き状態関係」を考案し，上記のプロセスに組み込む。まず，開始時刻で昇順ソートされた状態系列 S において，2 つの状態 $\{(S_i, S_j) : i < j\}$ の時間差 d_{ij} を次式により定義する。

$$d_{ij} = b_j - e_i \tag{1}$$

ここで，b_j は状態 S_j の開始日時，e_i は状態 S_i の終了日時である。S_j は S_i より後にあるため，$d_{ij} > 0$ の場合は before 関係，$d_{ij} \leq 0$ の場合は co-occur 関係となる。次に，d_{ij} のレンジに応じて，表 1 に示す 6 種類の期間付き状態関係を定義する。状態の数及び期間の定義は任意であるが，ここでは日，週，月の単位で 3 段階で抽象化した。これらの関係においても同様に時間制約があり，関係マトリクスにおいて，いずれかの before 関係が先行する場合，それ以降，同じ行では before 関係のみが許容される。また，$bd < bw < bm$ の関係性から，後ろに来る before 関係は前の before 関係と同等以上の長さである場合のみ許容される。同様に，$cd < cw < cm$ の関係性から，後ろに来る co-occur 関係は，前の co-occur 関係と同等以下の長さである場合のみ許容される。これらの制約に基づき，不要な候補パターンを除外する。その他の手続きは基本的に元のアルゴリズムと同じである。

3. 実験結果

提案法をテストデータ及び医療データに適用し，基礎的な性質を調べた。プログラムは Batal らの論文における Algorithm 1 および Algorithm 2 を Python により独自に実装したものに，前節で提案した 6 種類の期間付き状態関係を取り扱うための拡張を加えて使用した。実験で用いたパラメータは，クラス内最小サポート値 $\sigma_y = 0.10$，MPTP を判別するための二項検定の片側有意水準は論文と同じく $\alpha = 0.01$ とした。実験はワークステーション (Xeon X5672 3.2GHz 1P/4C, 8GB Mem, SSD, CentOS 6.8) 上で実施した。

3.1 テストデータ

検査項目 1 が低値 (V1=L)，項目 2 が高値 (V2=H)，薬剤 1 を投与 (M1=ON) という 3 つの状態からなる簡素な系列をベースとして，状態間の関係とデータ数を様々に変化させたデータセットを人工的に生成し，MPTP の抽出実験及び所要時間の評価を行った。図 3 にテスト系列の構成を示す。状態の出現順序は早いものから順に V1=L，V2=H，M1=ON と固定し，その継続間隔と出現間隔を日，週，月の単位で変化させた 5 種類のパターンを構成した。例えばタイプ 1 は V1=L が基準日 (Day=0) から 1 日間，V2=H が約 2 週間後となる 17 日目から 1 日間，さらに M1=ON が約 1ヶ月後となる 35 日目から 1 日間生じるもので，2 種類の期間付き関係 $\langle bw \rangle$(several weeks before) と $\langle bm \rangle$(several months before) を用いて同図中欄に示すパターンとして記述される。状態の出現順を同一に固定していることから，状態系列はタイプに関わらず <V1=L,V2=H,M1=ON> となり，関係マトリクスのみタイプごとに異なるパターンとなる。正例群 (class P) 及び負例群 (class N) における各タイプの配分は同図右欄に記載のとおりである。このテストデータに含まれる MPTP は，タイプ 1 に相当する 3-pattern が 1 つと，2-pattern が 3 つ (V1=L $\langle bw \rangle$ V2=H，V1=L $\langle bm \rangle$ M1=ON，V2=H $\langle bw \rangle$ M1=ON) の計 4 つである。

データ数 n を 100 から 10,000 まで 4 段階に変化させ，MPTP 抽出に要した時間を計測した結果を表 2 に示す。本実験では，期間付き関係の増減による影響をあわせて評価するため，関係 before, co-occur それぞれが 1 種類 (B1C1)，2 種類 (B2C2)，3 種類 (B3C3) である場合の 3 条件で計測を行った。B1C1 は表 1 における bd, bw, bm を単一の関係に，cd, cw, cm を単一の関係に集約したもので，従来法とほぼ等価である。B2C2 は bd と bw，cd と cw をそれぞれ単一の関係に集約したものである。所要時間は次の 2 段階に分けて測定した。なお，所要時間はいずれも測定 5 回の平均値 ± 標準偏差を示している。

1. データ読み込み後，期間付き関係に基づいて全てのケースを状態系列へ変換するプロセス (同表第 3 列)

2. 1. の終了後，期間付き関係を用いて頻出 1-pattern から順次長い候補パターンを生成し，1. と照合して全ての頻出パターン及び MPTP を抽出するプロセス (同表第 4 列)

同表から，1. については n の増加に伴い大きく増加する傾向が見られたが，関係の種類数による違いは顕著には見られなかった。関係の種類が増えることで，状態間の時間差から対応する期間付き状態関係を判別する選択肢が増加するが，1 ケースの系列長が高々3 であり，含まれる関係の数が 3 つと少ないことから，全体の処理時間への影響は表出していないと考えられる。一方，2. については n に加えて関係の種類数による差異が見られる。長さ k の頻出パターンから生成される長さ $k+1$ の候補パターンの数は，関係の組み合わせの数に比例することから k の増大とともに急激に多くなるが，生成された候補パターンと入力系列との照合処理は，元となる長さ k の頻出サブパターンの適合事例集合に対してのみ行われるため，枝刈りの効果によって増加が抑制される。同図第 5 列に，候補パターンと入力系列の照合が行われた回数を示す。本実験のテストデータを構成する 3 つの状態 (V1=L，V2=H，M1=ON) は，いずれも頻出 1-pattern である。そこから生成される候補 2-pattern は，例えば B1C1 の場合，関係が 2 種類であることから，3 状態× 2 関係× 2 状態 (両側が同じ状態は省かれる)=12 個である。これに n をかけた数が 2-pattern の照合回数であり，$n = 100$ の場合，1,300 回中 1,200 回が 2-pattern の照合，残る 100 回が 3-pattern の照合である。B3C3 の場合，関係が 6 種類であることから，候補 2-pattern の数は $3 \times 6 \times 2$=36 個と 3 倍になり，2-pattern の照合回数も B1C1 と比べて 3 倍に増える。3-pattern の照合も 270 回と 3 倍近くになるが，もともと照合対象の系列が頻出 2-サブパターンの適合事例に限られるため，総数への影響は小さく抑制される。2. の所要時間の違いはこの性質を反映しているものと考えられる。

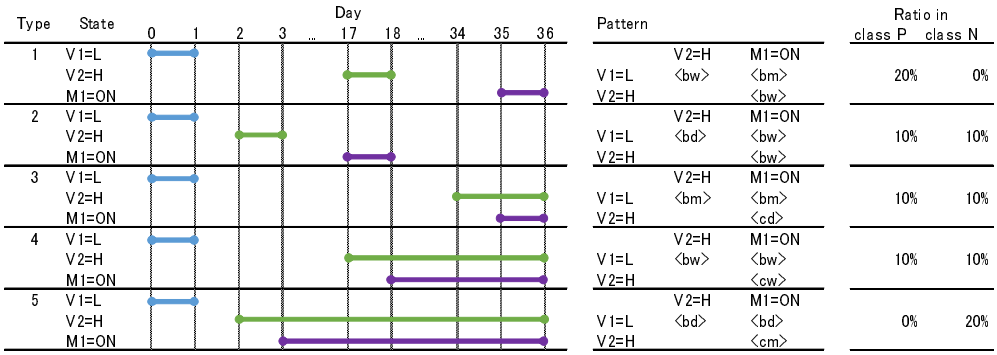

図 3: 5 種類のテスト系列の構成. 左：各タイプにおける 3 状態 (V1=H, V2=L, M1=ON) の時間関係. 中央：パターン. 右：クラス P 及び N における各パターンの配分割合.

表 2: テスト系列における MPTP 抽出の所要時間.

type of relations	number of cases	1. time for making state sequences [sec]	2. time for extracting MPTPs [sec]	number of sequences compared
B1C1	100	0.02±0.00	0.01±0.00	1,300
	1,000	0.64±0.00	0.04±0.00	13,000
	5,000	13.31±0.00	0.18±0.00	65,000
	10,000	67.34±0.27	0.37±0.00	130,000
B2C2	100	0.02±0.00	0.02±0.00	2,520
	1,000	0.64±0.00	0.07±0.00	25,200
	5,000	13.33±0.02	0.29±0.00	126,000
	10,000	67.22±0.07	0.61±0.00	252,000
B3C3	100	0.02±0.00	0.03±0.00	3,870
	1,000	0.64±0.00	0.10±0.00	38,700
	5,000	13.35±0.01	0.42±0.00	193,500
	10,000	67.35±0.13	0.88±0.01	387,000

3.2 医療データ

慢性関節リウマチを対象疾患として，検査・投薬の時系列と疾患活動性の変化を関連づけるパターン抽出実験を行った。クラス分類にはリウマチの疾患活動性の評価指標の一つである SDAI(Simplified Disease Activity Index) を使用し，2 年の観察期間を前半 1 年間と後半 1 年間に区分して，前半と後半で活動性 (SDAI の値を 4 段階に区分したもの) の最大値が減少した群を減少群 ($n = 102$)，不変であった群を不変群 ($n = 181$)，増加した群を増加群 ($n = 64$) とした。条件属性は上記 2 年間に実施された臨床検査の結果と，処方及び注射の実施歴である。なお，臨床検査は，結果が基準範囲外（H または L）であったものを対象として，H または L のラベルを使用した。

状態系列の長さの分布を図 4 に示す。右に裾の長い分布であり，中央値は 87，四分位範囲は 47-127，最大値は 761 である。テストデータの場合と異なり，$n = 347$ と比較的小さい一方，状態系列がかなり長いデータである。

以降では紙面の都合上，活動性減少群 ($n = 102$) に関する結果に絞って記載する。まず，同群における MPTP 抽出の所要時間の計測結果を表 3 に示す。各欄の内容は表 2 と同様であるが，所要時間についてはいずれも測定 3 回の平均値 ± 標準偏差である。また，可読性の観点から MPTP の長さの上限を $k = 4$ とした。

テストデータの場合と比べ，1. に対する 2. の割合が非常に大きい。このデータには計 95 個の頻出 1-pattern が存在しており，$k + 1$ 候補パターン生成の組み合わせの数が大きくなっていることが原因と考えられる。B2C2 の MPTP 抽出時間は他の 2 つと比べて短時間であったが，全体としては関係数の増加に伴い処理時間が増大している。同表第 4 列に照合回数を示す。この回数の内，2-pattern の照合は B1C1 の場合で約 38.5 万回，B3C3 の場合で約 115.6 万回行われており，関係数が 3 倍に増えたことを直接的に反映している。一方で，3-pattern 以上の照合を含めた比較総数は同表のとおり B3C3 は B1C1 に比べて 1.5 倍程度の増加であり，全体の処理時間に対する関係数増加の影響は比較的抑制されている。候補が多様化する一方で個々の候補の支持度が低下し，結果として頻出パターンの数が減少することが影響していると考えられる。

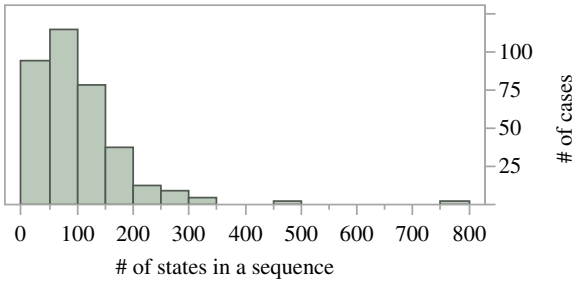

図 4: 全データ ($n = 347$) における状態系列の長さの分布.

表 3: 活動性減少群 (n=102) における MPTP 抽出の所要時間.

type of relations	1. time for making state sequences [sec]	2. time for extracting MPTPs [sec]	number of sequences compared
B1C1	10.58±0.26	1317.76±9.64	1,056,991
B2C2	12.41±0.23	1254.76±7.04	1,214,021
B3C3	14.74±0.31	1926.11±8.59	1,517,132

関係 B1C1(従来法に相当) 及び B3C3 により抽出した MPTP をそれぞれ表 4 及び表 5 に示す。パターンは confidence 値の降順で上位 8 個ずつ (B1C1 は全 16 個中，B3C3 は全 9 個中) の抜粋である。いずれにおいても抽出された MPTP に 4-pattern は無く，最長で 3-pattern であった。関係の種類の増加に伴い，B1C1 では MPTP であった一部のパターンが B3C3 では MPTP から外れている。一方で，B1C1(表 4) のパターン番号 2，3，8 などは B3C3(表 5) においても番号 1，2，3 として MPTP になっており，これらの状態関係については期間のばらつきが比較的小さいものと考えられる。提案法においては，例えば番号 1 の場合，[MMP3]=H → (数ヶ月) → [フォリアミン錠 5mg]=ON → (数ヶ月) → [好中球数]=L のように期間に関する情報がパターンに組み込まれ，従来法と比べてより詳しい関係性を表現可能となった。

表 4: 活動性減少群 ($n = 102$) で関係 B1C1(従来法に相当) により抽出された MPTP の例.

No	Conf	Supp	Pattern		
1	0.800	0.118	[ALP]=H	[ナトリウム (Na)]=L ⟨b⟩	
2	0.786	0.108	[MMP-3]=H [フォリアミン錠５ｍｇ]=ON	[フォリアミン錠５ｍｇ]=ON ⟨b⟩	[好中球数]=L ⟨b⟩ ⟨b⟩
3	0.737	0.137	[赤沈 1 時間値]=H [リウマトレックスカプセル２ｍｇ]=ON	[リウマトレックスカプセル２ｍｇ]=ON ⟨b⟩	[フォリアミン錠５ｍｇ]=ON ⟨b⟩ ⟨c⟩
4	0.733	0.108	[ナトリウム (Na)]=L	[血糖 (空腹時、随時）]=H ⟨c⟩	
5	0.733	0.108	[PDW]=L [血糖 (空腹時、随時）]=H	[血糖 (空腹時、随時）]=H ⟨c⟩	[好中球数]=H ⟨c⟩ ⟨c⟩
6	0.733	0.108	[血算-ヘマトクリット]=L [(リンパ数)]=L	[(リンパ数)]=L ⟨c⟩	[C 反応性蛋白 (CRP)]=H ⟨c⟩ ⟨b⟩
7	0.722	0.127	[血清アルブミン定量]=L [Lymph]=L	[Lymph]=L ⟨c⟩	[RDW-SD]=H ⟨c⟩ ⟨b⟩
8	0.684	0.127	[MMP-3]=H	[血算-白血球数]=L ⟨b⟩	

MMP-3: マトリックスメタロプロテイナーゼ-3, PDW: 血小板分布幅

表 5: 活動性減少群 ($n = 102$) で関係 B3C3 により抽出された MPTP の例.

No	Conf	Supp	Pattern		
1	0.846	0.108	[MMP-3]=H [フォリアミン錠５ｍｇ]=ON	[フォリアミン錠５ｍｇ]=ON ⟨bm⟩	[好中球数]=L ⟨bm⟩ ⟨bm⟩
2	0.706	0.118	[赤沈 1 時間値]=H [リウマトレックスカプセル２ｍｇ]=ON	[リウマトレックスカプセル２ｍｇ]=ON ⟨bm⟩	[フォリアミン錠５ｍｇ]=ON ⟨bm⟩ ⟨cd⟩
3	0.684	0.127	[MMP-3]=H	[血算-白血球数]=L ⟨bm⟩	
4	0.667	0.137	[カリウム (K)]=H	[ナトリウム (Na)]=L ⟨bm⟩	
5	0.667	0.118	[クロール (Cl)]=L		
6	0.606	0.196	[免疫グロブリン G(IgG)]=H		
7	0.611	0.108	[ケナコルト-A 筋注用 40mg/V]=ON	[総コレステロール (T-CHO)]=H ⟨bm⟩	
8	0.480	0.235	[Crea]=H		

MMP-3: マトリックスメタロプロテイナーゼ-3

4. おわりに

本稿では，Batal らの時系列マイニング法を拡張し，期間付き関係を導入することで状態の継続期間及び間隔をパターンへ組み込む方法を提案した。テストデータ及び医療データに対する適用実験では，関係の種類の増加に伴い候補パターンの組み合わせが増大し，特に低次のパターンの照合回数が大きく増加する一方，高次のパターンではサブパターンの支持度が低下し頻出パターンの数が減少することなどから全体の照合回数の増加は比較的抑制されていることが分かった。提案法により抽出された MPTP では，例えば状態間の間隔が数ヶ月の単位であるなど，時間的な関係性をより明確にパターンに組み込むことが可能になった。今回の使用した期間付き関係は，日，週，月を単位として抽象化するものであったが，a few, several 等の数量形容詞と組み合わせることで関係性を多様に表現することも考えられる。その場合の効率的なパターン比較や候補生成の方法についても今後検討を進めていきたい。

謝辞

本研究の一部は JSPS 科研費 (基盤研究 (C) 26330253) 及び AMED 臨床研究等 ICT 基盤構築研究事業 (医用知能情報システム基盤の研究開発) の助成による。

参考文献

[1] Batal I, Valizadegan H, Cooper GF, Hauskrecht M.; A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data. ACM Trans Intell Syst Technol. 4(4) (2013).

[2] Allen JF.: Maintaining knowledge about temporal intervals. In: Communications of the ACM. 26 (1983).