

## 科学研究費助成事業 研究成果報告書

平成 26 年 6 月 14 日現在

機関番号：15201

研究種目：基盤研究(C)

研究期間：2011～2013

課題番号：23500179

研究課題名(和文) 概念と時間を最適粒度で比較する時系列マイニング法の開発

研究課題名(英文) Development of a process mining system capable of handling temporal and compositional irregularities

研究代表者

平野 章二 (Hirano, Shoji)

島根大学・医学部・准教授

研究者番号：60333506

交付決定額(研究期間全体)：(直接経費) 4,000,000円、(間接経費) 1,200,000円

研究成果の概要(和文)：本研究では、時間的不均質性を有する質的な医療行為データを対象とし、適用頻度及び推移頻度からプロセスの「典型らしさ」を指標化、共通した診療プロセスの類型を生成する方法を開発した。耳鼻科系疾患158例に対する適用実験では、第2日を手術日とする4群間において、適用されたパスの有無及び種別をカテゴリとする例数分布に関して統計的に有意な差異が見られた。また同様に産科系疾患124例においても得られた群と各パスカテゴリの例数分布との間に有意な関連が見られた。これらにより、提案方法によって、手術等に関連した事前情報を与えることなく、典型らしさから診療プロセスの類似性を評価して類型化し得ることを示した。

研究成果の概要(英文)：In this research we developed a method for finding groups of common treatment processes based on the typicalness index, which reflects the global occurrence and transition frequencies of issued orders. We conducted experiments on otolaryngological disease data (158 cases) and observed statistically significant differences on the distributions of applied pathways, among the four clusters generated assessing the similarity of the typicalness indices of order sequences. Similarly, on obstetric disease data (124 cases), we observed statistically significant differences between the distribution of applied pathways and generated clusters. Based on these results, we demonstrated that by using the typicalness measure we could discover the groups of similar treatment processes without giving any a priori information about order types such as operations.

研究分野：総合領域

科研費の分科・細目：情報学・知能情報学

キーワード：知識発見とデータマイニング

1. 研究開始当初の背景

クリニカルパスは、診療計画を最適化しテンプレートとして共有するもので、医療の質的な標準化、在院日数の短縮等を図るツールとして注目を集めている。パスの作成方法は様々であるが、診療計画の妥当性に加えて各部門との連携を含む運用容易性を考慮する必要があり、白紙からの作成は医療従事者にとって非常に負担が大きい。そのため、実際には過去の治療例からスムーズに退院できた典型例を選び出し、その診療プロセスをたたき台として医療チームが検討を重ねパス化する等の手法がとられる。しかしながら、このような手法においてもなお、ひな型となる事例の抽出は症例の横断比較を伴う負担の大きい作業であり、計算機による支援が望まれている。

過去の症例から典型的な診療プロセスを模索する過程は質的時系列集合からのパターンマイニングといえるが、診療プロセスがある程度定型化されている疾患においても、実際に発出されるオーダ（医療者から発せられる処置、処方、検査、看護などの指示データ）には患者状態や検査部門の状況等に応じて大小様々な差異が生じ、結果として時間的にも内容的にも不均質な系列の集まりとなる。それらを念頭に、細かな差異を吸収しつつ、基本的なプロセスとしての共通性やまとまりを捉える手法の開発が課題であった。

2. 研究の目的

本研究では、不均質さを伴う質的プロセス系列の比較分類法について研究を行う。具体的な応用事例/実験対象として前述の時系列診療記録の横断的分析を取り挙げ、病院情報システムに蓄積された実際のオーダ歴から、クリニカルパスの骨格となり得るレベルの診療行為の類型（適用率の高い典型的なオーダからなる診療プロセス）を獲得するシステムの構築を目指す。

3. 研究の方法

提案方法では、まず全ての事例の情報を元に、各フェーズ（段階）において各オーダ項目がどの程度の頻度で適用されているか（適用頻度）、また、隣接フェーズ間でどのオーダ組が接続関係として多用されているか（推移頻度）を集計する。次に、これらの集計情報をもとに、各事例のオーダ系列の「典型らしさ」を評価する。さらに、この典型らしさ指標をもとに事例をクラスタリングし、類似した事例からなるグループを構成する。最後に、各グループにおいて指標値の高い事例のオーダ系列を典型的なプロセスとして抽出する。以上の手続きにより、適用率の低いオーダに関する差異を抑えつつ、データに複数のプロセスが内包される場合もそれぞれを表出させる。

3.1 適用頻度及び推移頻度の算出

まず、全ての事例のデータを用いて、各フェーズにおけるオーダの適用頻度を算出する。フェーズ数を $P$ とし、フェーズ $p \in P$ にお

いて、コード $c_k$ のオーダが適用される頻度を $Freq(c_k^{[p]})$ とすると、

$Freq(c_k^{[p]}) = \left\| \left\{ x_{im}^{[p]} \in U \mid code(x_{im}^{[p]}) = c_k \right\} \right\|$  と定義する。ここで、 $U = \{x_1, x_2, \dots, x_n\}$  は全事例のオーダ集合を、 $code(x_{im}^{[p]})$  は事例 $x_i$ における $m$ 番目のオーダ $x_{im}^{[p]}$ のオーダコードをそれぞれ表す。また、記号 $\|\cdot\|$ はこの集合に含まれる要素の数を示す。上式は、フェーズ $p$ において、コード $c_k$ のオーダを有する事例を数え上げることに相当する。

次に、隣接する2つのフェーズにおけるオーダの推移頻度を算出する。フェーズ $p$ においてコード $c_k^{[p]}$ が、フェーズ $p+1$ においてコード $c_l^{[p+1]}$ が、いずれも同一事例にて出現する頻度を $Freq(c_k^{[p]}, c_l^{[p+1]})$ とすると、

$Freq(c_k^{[p]}, c_l^{[p+1]}) = \left\| \left\{ (x_{im}^{[p]}, x_{in}^{[p+1]}) \in U \mid (code(x_{im}^{[p]}) = c_k) \wedge (code(x_{in}^{[p+1]}) = c_l) \right\} \right\|$  と定義する。上式は、隣接フェーズにおける2つのオーダの共起頻度を求めることに相当する。

次に、隣接する2つのフェーズにおけるオーダの推移頻度を算出する。フェーズ $p$ においてコード $c_k^{[p]}$ が、フェーズ $p+1$ においてコード $c_l^{[p+1]}$ が、いずれも同一事例にて出現する頻度を $Freq(c_k^{[p]}, c_l^{[p+1]})$ とすると、

次に、隣接する2つのフェーズにおけるオーダの推移頻度を算出する。フェーズ $p$ においてコード $c_k^{[p]}$ が、フェーズ $p+1$ においてコード $c_l^{[p+1]}$ が、いずれも同一事例にて出現する頻度を $Freq(c_k^{[p]}, c_l^{[p+1]})$ とすると、

図1に適用頻度及び推移頻度の表現方法を図示する。データセットに含まれるオーダコードの総数を $C$ とすると、適用頻度は同図(a)に示すように $C \times P$ の行列として表現でき、その要素が $Freq(c_k^{[p]})$ である。また、推移頻度に関しては同図(b)に示すように一組の隣接フェーズ $(p, p+1)$ に対して $C \times C$ の行列が一つ作成され、その要素が $Freq(c_k^{[p]}, c_l^{[p+1]})$ である。推移頻度の行列はフェーズ数を $P$ とすると $P-1$ 個生成される。なお、オーダコードの集合 $\{c_1, c_2, \dots, c_C\}$ には、データセットに含まれる全てのオーダコードに加えて、「オーダ無し」を表す任意のコードを1つ含むものとする。

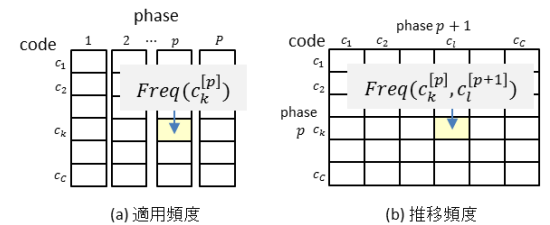


図1 適用頻度行列と推移頻度行列

3.2 「典型らしさ」の評価

前節で生成した適用頻度の行列と推移頻度の行列を用いて、各事例のオーダ系列の典型らしさを評価する。この評価は、フェーズ $p$ から $p+1$ を見る正方向(forward)と、フェーズ $p+1$ から $p$ を見る逆方向(reverse)の双方向で行う。図2に例を示す。同図において、実線枠で囲んだオーダはこの事例に含まれるオーダ項目を、点線枠で囲んだオーダはこの事例に含まれない(データセット中の他の事例に含まれる)オーダ項目を示す。また、オーダ枠の隣に添記した数値はそのオーダの適用頻度を、オーダを結ぶ直線上の数値は当該オーダ間の推移(共起)の発生頻度を示し、それぞれ前節で求めた $Freq(c_k^{[p]})$ 、 $Freq(c_k^{[p]}, c_l^{[p+1]})$ と対応している。例えばこの事例はフェーズ $p$ にオーダ $c_1^{[p]}$ を含んでおり、同様に $c_1^{[p]}$ を含む事例がデータセット全体で

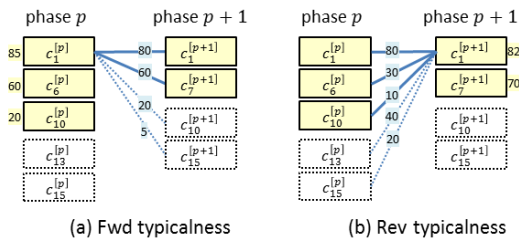


図2 典型らしさ指標

85件 (=  $Freq(c_1^{[p]})$ ) がある。いま、同図(a)に示す正方向の関係を考える。フェーズ  $p$  にオーダ  $c_1^{[p]}$  を含む事例の内、 $p+1$  においてオーダ  $c_1^{[p+1]}$  を含むものはデータセット全体で80件あり、 $c_7^{[p+1]}$ ,  $c_{10}^{[p+1]}$ ,  $c_{15}^{[p+1]}$  を含むものはそれぞれ全体で60件、20件、5件ある。データセット全体でフェーズ  $p+1$  に位置するオーダはこれら4種類であり、そのうち今回の事例に含まれるものは実線枠で囲んだ  $c_1^{[p+1]}$ ,  $c_7^{[p+1]}$  の2つである。残りの2種類 ( $c_{10}^{[p+1]}$ ,  $c_{15}^{[p+1]}$ ) はこの事例に含まれないが、それらの頻度は20件、5件と低く、典型らしさという観点での寄与は小さいと考えられる。そこで、この事例に含まれるオーダの頻度和 ( $80+60=140$ ) を、全体の頻度和 ( $80+60+20+5=165$ ) で除すことで、接続して出現する頻度の高いオーダがこの事例にどの程度含まれているかを求める。その値を、元となるオーダ  $c_1^{[p]}$  の頻度 (=85) で重み付けすることで、 $c_1^{[p]}$  に関する正方向の典型らしさ指標 ( $85 \times 140 / 165 = 72.1$ ) を得る。同様の処理をフェーズ  $p$  の各オーダについて行い、総和をとってフェーズ  $p \rightarrow p+1$  の正方向の典型らしさ指標とする。以上を次式により定式化する。

$$FwdTypicalness^{[p]}(X_i) = \sum_{m=1}^{N^{[p]}} \left\{ Freq(c_m^{[p]}) \times \frac{\sum_{n=1}^{N^{[p+1]}} Freq(c_m^{[p]}, c_n^{[p+1]})}{\sum_{k=1}^{N^{[p+1]}} Freq(c_k^{[p]}, c_n^{[p+1]})} \right\}$$

ここで、簡単のためにオーダ  $x_{im}^{[p]}$  のコード  $code(x_{im}^{[p]})$  を  $c_m^{[p]}$  と記している。FwdTypicalness は、後ろに接続すべき頻度の高いオーダが欠落しているとき低い値をとる。同図(b)に示す逆方向についても同様に、

$$RevTypicalness^{[p]}(X_i) = \sum_{n=1}^{N^{[p+1]}} \left\{ Freq(c_n^{[p+1]}) \times \frac{\sum_{m=1}^{N^{[p]}} Freq(c_m^{[p]}, c_n^{[p+1]})}{\sum_{k=1}^{N^{[p]}} Freq(c_k^{[p]}, c_n^{[p+1]})} \right\}$$

と定義する。RevTypicalness は、前段に接続すべき頻度の高いオーダが欠落しているとき低い値をとる。これらの指標はいずれもオーダ数の影響を受けるため、次式のとおり正逆両方向の指標の和をオーダ数で平均し、これを事例  $X_i$  に関するフェーズ  $p$  の典型らしさ指標とする。

$$Typicalness^{[p]}(X_i) = \frac{FwdTypicalness^{[p]}(X_i) + RevTypicalness^{[p]}(X_i)}{N^{[p]} + N^{[p+1]}}$$

事例  $X_i$  のオーダ系列全体の典型らしさは、全てのフェーズについて指標値とオーダ数を積算し平均をとることで次式のごとく得られる。

$$Typicalness(X_i) = \frac{\sum_{p=1}^P \{FwdTypicalness^{[p]}(X_i) + RevTypicalness^{[p]}(X_i)\}}{\sum_{p=1}^P (N^{[p]} + N^{[p+1]})}$$

### 3.3 典型らしさに基づくクラスタリング

前節において定義した典型らしさ指標を用いて事例のクラスタリングを行う。ここでは、指標をフェーズ単位で比較することにより、単に合計値の類似性ではなく、指標の分布の類似性を反映させる。事例  $X_i$  と  $X_j$  の相違度  $Dissim(X_i, X_j)$  を次式により定義する。

$$Dissim(X_i, X_j) = \sum_{p=1}^P \{Typicalness^{[p]}(X_i) - Typicalness^{[p]}(X_j)\}^2$$

なお、事例  $X_i$  と  $X_j$  でフェーズの数が異なる場合、存在しないフェーズの typicalness を0として取り扱う。例えば、 $X_i$  が8フェーズ、 $X_j$  が9フェーズで構成されるとき、9フェーズ目の相違度は  $X_i$  側を0として算出する。

## 4. 研究成果

島根大学病院の病院情報システムから抽出したオーダデータに対して提案手法を適用し、どのようなクラスタ/典型事例が得られるかを調べた。ここでは、複数種類のクリニカルパスが既に作成され運用されている耳鼻咽喉科系と産科系の2疾患を対象として、オーダの典型らしさの類似性に基づき構成される事例群と、クリニカルパスの適用有無あるいは適用されたクリニカルパスの種類との関係を評価した。

実験では、提案法により各事例の典型らしさ指標及び事例間の相違度を算出し、得られた相違度行列をもとに階層的クラスタリング法の一つであるWard法を適用してクラスタ分析を実施した。

### 4.1 耳鼻科系疾患に対する適用結果

この疾患においては、ある部位Aの摘出術に関する3種類のクリニカルパスが作成されており、本稿ではそれぞれ(1)部位A摘出術、(2)部位A摘出術(AM手術用)、(3)部位A摘出術(PM手術用)と呼称する。このうち、(1)が最初に作成され、それが後に(2)の午前手術用と(3)の午後手術用に分岐した。いずれのパスも複数回の改訂が行われ、(1)は2つ、(2)は5つ、(3)は5つの版が混在している。今回抽出した158例のうち、パスが適用されたものは78例、非適用は80件であった。

Ward法により生成された樹状図を図3に示す。大きく分けて2つないし3つの群が構成されているが、ここでは過度の併合を避け、

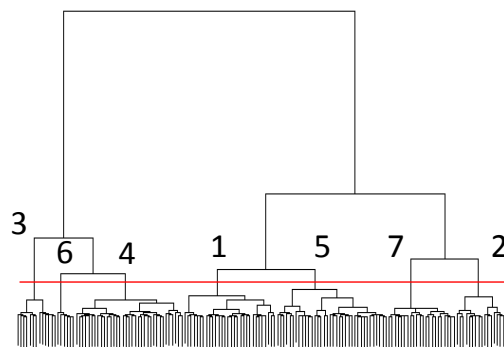


図3 典型らしさ指標に基づく樹状図 (耳鼻科系疾患)

比較的早期に相違度のステップが大きくなる7クラスタで分類し各クラスタの特徴を観察した。以後の参照のため、各クラスタの番号を樹状図上に添記した。

各クラスタごとに、例数、手術フェーズ(手術オーダが何日目にあるか)の平均値と標準偏差、典型らしさ指標の平均値と標準偏差、パスの適用例数、適用率をそれぞれ集計したものを表1及び表2に示す。表におけるクラスタの記載順は樹状図における記載順と対応している。

表1 各クラスタの例数, 手術フェーズ, 典型らしさ指標 (耳鼻科系疾患)

Cluster index	Number of cases	Phase of operation (Mean ± SD)	Typicalness index (Mean ± SD)
3	13	3.9 ± 0.3	17.7 ± 3.8
6	6	-	6.9 ± 5.6
4	35	4.0 ± 0.7	12.5 ± 3.5
1	30	2.0 ± 0.0	30.9 ± 4.5
5	36	2.0 ± 0.2	24.1 ± 4.9
7	22	2.2 ± 0.6	21.9 ± 3.6
2	16	2.0 ± 0.0	23.6 ± 6.1
Total	158	-	21.3 ± 8.1

表2 各クラスタのパス種別例数とパス適用率 (耳鼻科系疾患)

Cluster index	Number of cases for each type of pathway				Total	Ratio of applying pathways
	Region A extraction	AM OP	PM OP	No pathways applied		
3	1	0	2	10	13	0.23
6	0	0	0	6	6	0.00
4	5	3	2	25	35	0.29
1	4	10	14	2	30	0.93
5	6	4	9	17	36	0.53
7	2	3	2	15	22	0.32
2	1	7	3	5	16	0.69
Total	19	27	32	80	158	0.49

クラスタ3, 6, 4においては、いずれもパス適用率が3割未満であり、特にクラスタ6はパス非適用の事例のみから構成されている。また、これらのクラスタにおける手術日フェーズは平均が第4日周辺にあり、第2日周辺を平均とするクラスタ1以下の群と明瞭な違いが観察された。典型らしさ指標はクラスタ6が最も低く、他の2つのクラスタも同様に低い。これらから、図4の樹状図における最初の大分岐は、手術日の違いに起因する診療プロセスの違いを反映しており、左側の枝には比較的典型的でない事例が分類されているものと推察される。なお、クラスタ6においては、6例中4例で手術オーダが発出されていないため、手術日フェーズの平均値等は算出していないが、残り2例の手術日はそれぞれ第4日及び第7日であった。

一方、樹状図で右側の枝に位置するクラスタ1, 5, 7, 2を見ると、手術日フェーズの平均はいずれも第2日前後にあり、また典型らしさ指標も高い。特にクラスタ1はパス適用率が9割を超えており、典型らしさ指標も最も高いことから、パスに基づく比較的定型的なプロセスが適用された事例が集合していると言える。これら4つのクラスタは、樹状図上の第2分岐において(1, 5)(7, 2)の2群を構成する。表2におけるパス名称ごとの例数を見ると、クラスタ1, 5においては部位A摘出術(PM手術用)の割合が他と比べて高く、逆にクラスタ7, 2において

は部位A摘出術(AM手術用)の割合が他と比べて高い。これら4つのクラスタ(1, 5, 7, 2)におけるパス種別(非適用を含む4種類)の例数分布には有意な差異( $p < 0.001$ , Fisher's exact test)がみられた。パス非適用の事例も多く、違いは必ずしも明確ではないが、手術の時間帯によるプロセスの違いが典型らしさ指標に反映され、群として表出したものと考えられる。

#### 4.2 産科系疾患に対する適用結果

この疾患においては初産と経産で区分される2種類のパスが使用されている。今回抽出した124例のうち、パスが適用されたものが25例、非適用は99件であった。なお、この疾患は手術を伴わないものである。

図4にWard法により生成された樹状図を示す。耳鼻科系疾患の場合と同様に、ここでは4クラスタに区分して特徴を分析した。表3及び表4に各クラスタの例数、典型らしさ指標、パスの適用例数及び適用率をそれぞれ示す。典型らしさ指標はクラスタ4と2で比較的高値であり、最も多くの事例を含むクラスタ1とクラスタ3は低値であった。この4群における適用パスの分布には非常に明瞭な違いが見られ、クラスタ1はほぼ全てパス非適用の事例で構成される一方、クラスタ2は経産婦向けパスの適用事例が集中した。また、初産婦向けパスの適用事例はクラスタ4に集中した。これら4群におけるパス種別(非適用を含む3種類)の分布には、有意な差異( $p < 0.001$ , Fisher's exact test)がみられた。

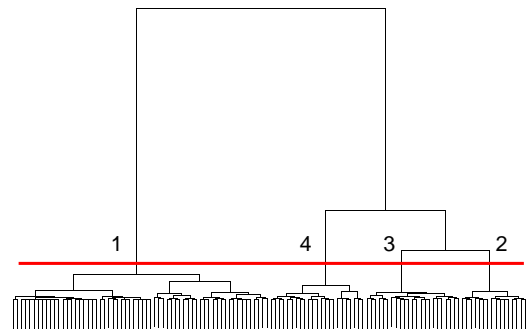


図4 典型らしさ指標に基づく樹状図 (産科系疾患)

表3 各クラスタの例数, 手術フェーズ, 典型らしさ指標 (産科系疾患)

Cluster index	Number of cases	Typicalness index (Mean ± SD)
1	62	5.2 ± 2.1
4	23	13.1 ± 2.1
3	23	7.9 ± 2.2
2	16	12.1 ± 2.3
Total	124	8.0 ± 4.0

表4 各クラスタのパス種別例数とパス適用率 (産科系疾患)

Cluster index	Number of cases for each type of pathway				Ratio of applying pathways
	For multiparas	For primiparas	No pathways applied	Total	
1	1	0	61	62	0.02
4	0	9	14	23	0.39
3	0	1	22	23	0.04
2	14	0	2	16	0.88
Total	15	10	99	124	0.20

これらの結果は手術などオーダ種別に関する事前情報を一切用いずに得られたものであり、2疾患を対象とした限られた分析ではあるが、提案法によりプロセスの差異を指標化し、類似したプロセスからなる群を生成できたことを示すと考えられる。

#### 5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[雑誌論文] (計1件)

- ① Shoji Hirano and Shusaku Tsumoto: Clustering of non-metric proximity data based on bi-links with epsilon-indiscernibility. Journal of Intelligent Information Systems, Volume 41, Issue 1, pp 61-71, August 2013.

[学会発表] (計6件)

- ① Shoji Hirano and Shusaku Tsumoto: Cluster analysis of treatment processes based on the typicalness measure for building clinical pathways. 2013 IEEE BIBM Workshop on Biomedical and Health Informatics, 2013年12月19日, Shanghai, China.
- ② Shoji Hirano and Shusaku Tsumoto: Clustering of order sequences based on the typicalness index for finding clinical pathway candidates. 2013 IEEE ICDM Workshop on Data Mining for Services, 2013年12月7日, Dallas, USA.
- ③ 平野章二, 津本周作: 典型らしさ指標に基づくクリニカルパス候補生成の試み. 第14回日本医療情報学会学術大会, 2013年11月22日, 神戸ファッションマート, 神戸市.
- ④ Shoji Hirano and Shusaku Tsumoto: Mining clinical pathway candidates from order history based on the clustering of order sequences. 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013年10月16日, Manchester, UK.
- ⑤ 平野章二, 津本周作: オーダ時系列の分類に基づくクリニカルパス候補生成の試み. 2013年度人工知能学会全国大会, 2013年6月4日, 富山国際会議場, 富山市.
- ⑥ 平野章二, 津本周作: クリニカルパスの自動生成に向けたオーダデータの統計解析. 第93回人工知能学会知識ベースシステム研究会, 2011年8月9日, 島根大学, 出雲市.

[図書] (計0件)

[産業財産権]  
○出願状況 (計0件)

名称:  
発明者:  
権利者:  
種類:  
番号:  
出願年月日:  
国内外の別:

○取得状況 (計0件)

名称:  
発明者:  
権利者:  
種類:  
番号:  
取得年月日:  
国内外の別:

[その他]  
ホームページ等

#### 6. 研究組織

##### (1) 研究代表者

平野 章二 (HIRANO, Shoji)  
島根大学・医学部・准教授  
研究者番号: 60333506

##### (2) 研究分担者

( )

研究者番号:

##### (3) 連携研究者

( )

研究者番号: