2005－ＩＣＳ－140（６）
2005／8／1

依存関係に着目した系列パターン再構成

植野 研† 北原 洋一† 林 俊夫‡ 櫻井 茂明† 折原 良平†

† （株）東芝 研究開発センター 知識メディアラボラトリー 〒212-8582 川崎市幸区小向東芝町 1
‡ （株）東芝 東京都港区芝浦 1-1-1

E-mail: † {ken.ueno, youichi.kitahara, shigeaki.sakurai, ryohei.orihara}@toshiba.co.jp,
‡ toshio7.hayashi@ toshiba.co.jp

あらまし 系列パターンマイニングによって時系列データから複数の頻出系列パターンを生成し，パターン間の依存関係と推移の非対称性を使って特徴的な頻出系列パターンを再構成する方法を提案する．系列パターンを依存関係に着目して結合させると，すべての時間ずれの組合せを包含させることができる．定期健康診断データを用いた検証実験において，本方法により，この結合系列パターンと同期系列パターンとを推移頻度の対称性に基づく基準で比較することで，時間ずれを含む特徴的な系列パターンを抽出することが可能であり，特徴的な系列パターンを再構成させることが可能であることが分かった．

キーワード 系列パターンマイニング，依存関係，パターン結合，時系列イベントパターンマイニング，定期健康診断データ解析

Reconstruction of Sequential Patterns based on Dependencies

Ken Ueno† Youichi Kitahara† Toshio Hayashi‡ Shigeaki Sakurai† Ryohei Orihara†

† Knowledge Media Lab. Corporate Research and Development Center, Toshiba Corporation 1 Komukai-toshiba-cho, Saiwai-ku, Kawasaki-shi, Kanagawa, 212-8582 Japan
‡ 1-1-1 Shibaura, Minato-ku, Tokyo, Japan

E-mail: † {ken.ueno, youichi.kitahara, shigeaki.sakurai, ryohei.orihara}@toshiba.co.jp,
‡ toshio7.hayashi@ toshiba.co.jp

Abstract This paper proposes a pattern reconstruction method to acquire characteristic sequential patterns based on dependencies. The patterns which are joined by their dependencies can include any time lags. To compare the joined patterns with simultaneous patterns, the characteristic patterns with time lags can be extracted by c-ratio proposed in this research. With the pattern reconstruction method, we found that characteristic patterns are successfully induced from our health check database with the reconstruction method based on dependencies.

Keyword Sequential Pattern Mining, Dependency, Pattern Join, Time Series Event Pattern Mining, Health Check Data Analysis

1. はじめに

1.1. 背景

系列データからのマイニングに関する研究が近年注目されている．その応用領域は，医療時系列データの解析[7]，テキストデータからのイベント変化の発見[3,4]，多変量波形データからの知識発見[2]など広範にわたっている．しかしながら，企業における定期健康診断データからの生活習慣・検査値の解析には，仮説検証型の共分散分析やロジスティック回帰などの統計解析手法を適用する場合が多く，仮説の構造自身を探し出すために系列パターンマイニングを利用した解析はこれまで試みがなされてこなかった．

日本人の疾患の中で，高血圧はかなりの割合を占める．高血圧は，二次性高血圧（腎臓疾患・内分泌疾患などによるもの）よりも，多因的な本態性高血圧（食生活習慣，メンタル，遺伝などによるもの）が大半である．しかしながら，生活習慣と本態性高血圧の関係はいまだ研究の余地がある．また，定期健康診断において生活習慣に関するデータはおもに問診表によって取得されるが，その問診項目は通常非常に多数に及ぶためこれらの中から血圧に影響を与えている複合要因を見極めるのは困難である．さらに各受診者の数年から数十年に渡る生活習慣の継続・変化が現在の検査値に影響を与えることも考えられるが複合要因と継続・

変化を同時に解析する試みはあまり見られなかった．

1.2. 解析上の問題

血圧に影響を与える生活習慣の複合要因に関しては，複雑な解析が必要となる．統計分野で用いられている共分散分析，因子分析，変数選択を伴う重回帰分析，共分散構造分析などの方法を用いて，血圧と生活習慣の関連性に関する多くの研究がある[6]．しかしながら，解析に先立って交絡要因を見極めることが難しいという問題がある．また，これらの研究では，時間変化についての考察が少ないことも課題のひとつとして挙げられる．実際に，医療職（産業医，看護師，保健師）からは時間的な推移を考慮した要因解析が求められている．たとえば，塩分摂取が血圧を上昇させることは，十分にデザインされた実験条件の下での実験により医学分野ですでに明らかになっているが，数年単位での実際の生活の中での影響を考える場合は，継続的な食習慣がどれくらいの時間で血圧に影響を及ぼしているのかが重要な要件となってくる．このような時間的遅延を含めた解析は複雑な構造を扱わなければならない問題があった．

さらに，医療職からは，大規模な定期健康診断データの解析結果を健康指導に直接生かすことで，受診者の健康増進のモチベーションを高めることが求められている．指導上のニーズとして，特に，解析結果の可読性が求められていると考えられる．このような問題意識から，医療職，情報管理部門などの大規模な協力体制の下で，２００４年度から個人情報保護法を考慮した上でデータ管理・活用体制の整備を開始した．医療職の助言のもとでいくつかの候補要因を選択し，定期健康診断データの検査値・問診回答データからの時系列複合要因解析に当たった．

時系列複合要因解析の予備実験において，当初，AprioriAll[1,5]をベースとして制約パターンならびに系列 ID 集合を記録できるように改良した系列パターンマイニングエンジン[3]を実装して解析を行っていた．解析する中で，時系列複合要因に内在する時間ずれをどのように取り扱うかの問題が浮上していた[4]．

2. 目的

本研究では，以上の３点から，時間軸を考慮した複合要因を，制約パターン付きの系列パターンマイニングアルゴリズムで洗い出す方法を提案する．代表的な時系列パターンマイニングの方法では，最小支持度に基づく候補の枝刈りを用いるのが一般的である．しかしながら，ユーザがあらかじめ決定した最小支持度のみを使って枝刈りを行うと，複合的な要因が絡む場合に特徴的な系列パターンを取り逃してしまう問題がある．本稿では，複数の系列パターンの依存関係を考慮してパターン同士を連結させることで，時間ずれ[2]を許容した特長的な系列パターンを救い出し，系列パターンを再構成する方法を提案する．依存関係とは，２つ以上の頻出系列パターンにおいて，各系列パターンを満たしている系列データ ID 集合の積集合要素数で決まる関係のことで，この積集合要素数が大きいほど頻出系列パターン同士の依存関係が強いといえる．上記の方法を用いて，本稿では，実際の定期健康診断データにおける血圧と生活習慣との経時的共起推移パターンの解析例を示し，特徴的な複合系列パターンを求めることが可能であることを示す．

3. 定期健康診断データ

定期健康診断データには，検査値データと問診データの２つがある．検査値データには，血圧や中性脂肪などの計測値が定量的に記録されている．また，総合判定のために判定基準に基づく定性値が併記されている検査項目もある．一方，問診データには，生活習慣などに関する質問の回答が記録されている．食生活習慣に関する問診には，たとえば，「塩分の濃い食事を週に何回とりますか？」という問いに対して「１．毎日 ２．週に３・４回，３．ほとんどとらない」のうちいずれかを選択する項目や，「野菜をどれくらいとっていますか？」という問いに対して「１．毎日 ２．週に３・４回，３．ほとんどとらない」のうちいずれかを選択する項目などがある．

検査値データ，問診データはともに，各受診者の各項目に対する計測値（問診回答）から成るテーブルである．各テーブルには欠損値や不正値が微少ながら含まれている．後述する系列パターンマイニングアルゴリズムでは，欠損値があってもそのまま解析が可能であるため，欠損がある場合でもそのままデータを用いた．不正値に関してはパターン生成過程で除外した．検査値データ，問診データの各レコードは，「暗号化された受診者 ID」，「検査年」，「各項目の検査値（問診回答）」からなっている．本研究の目的は，生活習慣が血圧に及ぼす影響を調べることであるため，検査値データと問診データを結びつける必要がある．そこで，各データから必要のある部分だけを抽出し，暗号化された受診者 ID を用いてこれらの抽出された表をジョインして用いた．なお，個人情報保護の観点から，個人を特定できる情報は暗号化した上で解析にあたった．

4. 血圧値の離散化方法

定期健康診断データにおける血圧検査値は，収縮期血圧と拡張期血圧の２つに分かれている．今回使用した各血圧検査値は，G<問題なし>，Y<注意>，R<要精密検査>の３段階判定値に離散化され，受診者に提示

されている．しかしながら，時系列変化を見るためにはこの３段階判定では荒すぎてパターンの変化を捉えきれないことが予備実験で明らかとなった．この理由から３段階判定の各判定値を更に３等分して細分化し９段階判定に再離散化した（図１）．離散化の閾値は以下の式で決定した．

TH1 = 140, TH2 = 160, ST = 3,
⊿TH = (TH2 -TH1) / ST
NTH_1 = NTH2 - ⊿TH, NTH_2 = NTH_3 - ⊿TH
NTH_3 = TH_1
NTH_4 = TH_1 + ⊿TH, NTH_5 = TH_2 - ⊿TH
NTH_6 = TH_2
NTH_7 = NTH_6 + ⊿TH, NTH_8 = NTH_7 + ⊿TH

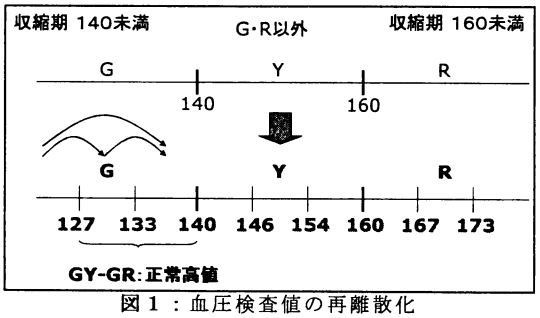

図１：血圧検査値の再離散化

拡張期も同様に再離散化した．収縮期の血圧判定値をJsys，拡張期の血圧判定値をJdiaとした．その後，以下の方法で９段階判定値BJを決定した.Gの場合はどちらか小さいほうをBJとし，条件を厳しくした．

BJ = min(Jsys, Jdia)　if Jsys == G? ∧ Jdia == G?,
BJ = max(Jsys, Jdia)　otherwise.

5. 系列パターンの生成方法

本研究では，制約パターンに基づく系列パターンマイニング法[3]を利用した．制約パターンに基づく系列パターンマイニング法を用いると，制約パターンと指定したパラメータに基づく共起系列パターンをすべて求めることができる．しかしながら，複数の要因が時間ずれを伴って共起する場合，すべての組合せについて制約パターンを記述しなければならない．また，時間ずれを含む個々の制約パターンは，最小支持率を下回ると枝刈りにより求めることができない．そこで，すべての時間ずれを包含するように複合要因の系列パターンを求める方法として，次章で「依存関係による系列パターンの再構成方法」を提案する．このようなすべての時間ずれを包含するように複合要因の系列パターンを**準同期結合パターン**(Quasi-synchronously joined patterns)と呼ぶことにする．また，すべてが同時に起こる推移パターンを**同期パターン**(Synchronously joined patterns)と呼ぶことにする．

6. 依存関係による系列パターンの再構成方法

以下の９ステップで頻出系列パターンを再構成させ，特徴的な頻出系列パターンを求める．

1．検査データから収縮期血圧値LBLと拡張期血圧値 HBL を暗号化済み受診者 ID，受診年度CYRとともに抽出する．

2．血圧検査値 LBL・HBL を再離散化する．RLBL・RHBLから血圧の総合判定を９段階血圧判定値BJとして算出する．

3．暗号化済み受診者 ID，受診年度 CYR をキーとして，抽出した問診表の必要項目を９段階血圧判定値 BJ とジョインする．暗号化済み受診者 ID，受診年度CYR，９段階血圧判定値，問診表項目からなるテーブル T が生成される．このテーブルは欠損値，外れ値を含む．欠損値には欠損マーカをつける．外れ値は頻度が低ければ系列パターンマイニングを適用する際にふるい落されるのでそのまま残す．

4．テーブル T を系列イベントデータ SD に変換する．系列イベントデータ SD は，各受診者の定期健康診断結果を年度ごとに並べなおした集合時系列イベントデータと言い換えることもできる．

5．系列イベントデータ SD の中から，あらかじめ設定した最小支持率 MinSup よりも上回る系列パターンを導き出して頻出系列パターン FP とする．なおここでは，制約パターン CP を導入し，系列パターンの各集合中の最大イベント数を規定した最大集合要素数 MaxSetElem をあらかじめ指定することで無駄な頻出系列パターンの生成を抑制させ，AからCまでの3つの頻出系列パターンを求めた.また，対応する系列イベントデータの ID 集合を各頻出系列パターンとともに記録した．

A) 「血圧」のみの推移
B) 「塩分の濃い食事をとる回数」の推移
C) 「野菜をとる回数」の推移
D) 「塩分の濃い食事をとる回数」が増加し，かつ，「野菜をとる回数」が減少する場合の「血圧判定」の同期結合パターンにより求めた推移

6．つぎに，ＡとＢの頻出系列パターンの依存関係を用いて系列パターンＡとＢの支持データＩＤ同士の積集合をとり，これらの系列パターンを結合させ，系列パターンＥとする．積集合の要素数を求める

ことでEの支持度を求める.

SupIDSet(E):= SupIDSet(A) ∩ SupIDSet(B)

Sup(E) := Freq(SupIDSet(E))

7. 先の系列パターンの場合と同様に結合系列パターンEの支持度を頻度マトリクスMにまとめる.頻度マトリクスMには系列パターンマイニングにより求まった頻出パターンに基づいて各血圧判定推移に当てはまった受診者数を記録する. 頻度マトリクスとは, 各パターンにより求まった血圧判定の推移に関するサポートカウントの表である. たとえば, 各生活習慣の条件下で血圧判定がGG→GYに推移したサポートカウントをMに入力する. パターンが出現しないものに関しては, 該当するマトリクスのセルは欠損値としておく. 同様の方法でF, Gを求める.

E) AとBより,「塩分の濃い食事をとる回数」が増加する場合の「血圧判定」の準同期結合パターンにより求めた推移

F) AとCより,「野菜をとる回数」が減少する場合の「血圧判定」の準同期結合パターンにより求めた推移

G) EとCより,「塩分の濃い食事をとる回数」が増加し, かつ,「野菜をとる回数」が減少する場合の「血圧判定」の準同期結合パターンにより求めた推移

8. 系列パターンGの頻度マトリクスと系列パターンDの頻度マトリクスとの推移頻度比をとり, 推移頻度比マトリクスMRを算出する.なお行列MRの各要素はmr_{ij}と表す. 比は$MR = M_G / M_D$で算出する. GとDとの比をとることで野菜・塩分の濃い食事とともに血圧が同時に変化する場合と, 野菜・塩分の濃い食とともに血圧が時間ずれを伴って変化する場合との頻度の比を取ることができる.

9. 基準値（たとえば平均値＋標準偏差値）を上回る比を持つ血圧判定の推移頻度のパターンをTPSetとし以下の式で頻出系列パターンを再構成する. ここで特徴的な推移パターンであるかどうかを判定する基準としてc-ratioを導入する.

$$c\text{-}ratio_{ij} = | mr_{ij} / mr_{ji} |$$

c-ratioは**推移の非対称性**を測る尺度として用いることができる. たとえば, ある条件下でGY→YGに着目した場合, GY→YG推移する頻度とYG→GYに推移する頻度の比をとることでGY→YGを特徴的な推移パターンと判定することができる. GY→YGとYG→GYとの頻度に差が認められる場合, 対称性が崩れ, GY→YGに特徴的な傾向があることが分かる. 予備実験では, 同期的に変化する多くの推移パターンがc-ratio≒1であることが分かっている. そこで本研究ではこのような推移パターンの対称性が崩れるものを特徴的なパターン集合ChPatとして取り出すためにc-ratio基準を用いた. ただし特徴的な比の最小閾値RTHは予備実験により RTH=1.5 とした. 特徴的なパターン集合ChPatは以下の方法で取り出した.

$$RTH = Average(MR) + stdev(MR)$$

$$TPSet := \bigcup_{ij} mr_{ij} \text{ if } mr_{ij} > RTH$$

Foreach $mr_{ij} \in TPSet$

Add mij to ChPat if $mr_{ij} - mr_{ji} > MinSup$

$\wedge$ $c\text{-}ratio_{ij} > RTH$

End

7. 実験

7.1. 実験設定

2000年から2004年度の社内定期健康診断データを複合推移パターン生成の実験対象とした. データ数は208,252件(58,654系列)である. それぞれ100以上の属性値をもつ検査・問診データから暗号化済受診者ID, 受診年度, 塩分, 野菜, 血圧の5属性を抽出した. 最小支持率はMinSup = 0.1 %とした.

7.2. 実験結果

はじめに, 血圧だけの推移パターンに基づく受診者数を図2に示す. ここから, ほとんどの受診者がGG内で変化無しであり, つぎにGY, GRの間での推移が多いことが分かった. また, Y以上の推移の頻度分布でも, 頻度のピークはYGであり, Gであれ Yであれ, 変化無しである頻度が高いことが分かった.

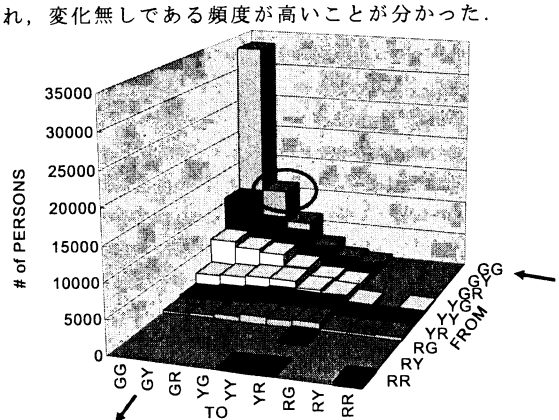

図2：血圧単独の推移パターン（z軸の人数は, 右

側の矢印の部分からはじめ，下の矢印の方向に推移する受診者数を示す．この矢印の例の場合，血圧判定がＧＧからＧＹに変わった受診者数は約 11,000 人であることが分かる）

つぎに塩分の濃い食事の問診回答が「無し」から「週に３・４日」に変化したときに同時に変化する血圧の推移パターンを求め，各推移で当てはまった受診者数をまとめると図３のようになる．

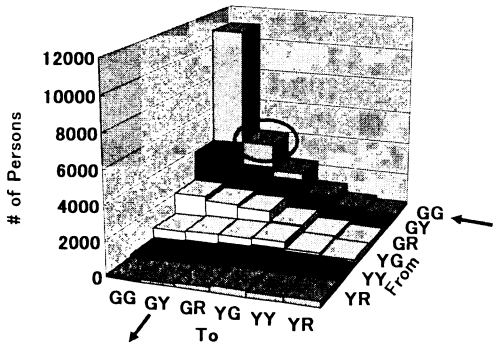

図３：塩分・血圧の準同期結合推移パターン（ｚ軸の人数は，右側の矢印の部分からはじめ，下の矢印の方向に推移する受診者数を示す．この矢印の例の場合，塩分の濃い食事が「無し」から「週３・４日」に変化したときに血圧判定がＧＧからＧＹに変わった受診者数は約 6,000 人であることが分かる）

野菜の摂取状況についても塩分と同様に，「毎日」から「週３・４日」に変化した場合の血圧判定の推移パターンを求めた．その結果，塩分の濃い食事の推移パターンとほとんど変わらないことが分かった．

ここで，塩分の濃い食事が「無し」から「週３・４日」へと変化し，同時に野菜の摂取が「毎日」から「週３・４日」へと変化したときの血圧の推移パターンを系列パターンマイニングにより求めた結果を図４に示す．最小支持率 0.1%の条件下では，ＧＧ，ＧＹ，ＧＲ間の推移パターンのみが求まった．また，推移マトリクスの受診者数は血圧推移の 1/20，塩分や野菜推移の 1/8 程度となった．頻度分布は血圧，塩分，野菜の頻度分布とほとんど変化が見られなかった．そこで，6 章にて示した「依存関係による系列パターンの再構成方法」を用いて，「血圧と塩分の系列パターン」と「野菜の系列パターン」を掛け合わせ，「血圧と塩分」「野菜」の準同期結合パターンを求めた．この準同期結合パターンに基づいて同様に推移マトリクスを示すと図５のようになった．

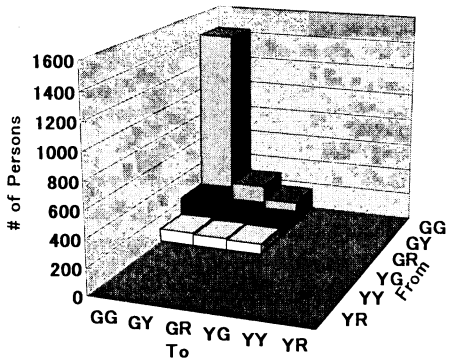

図４：塩分・野菜・血圧の同期結合推移パターン

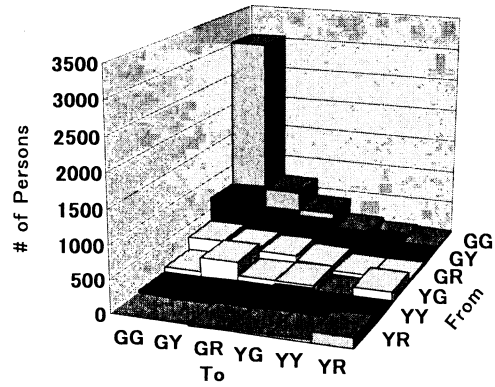

図５：塩分・野菜・血圧の準同期結合推移パターン

この準同期結合により，図４では求めることのできなかった推移パターンがＹＧ，ＹＹ，ＹＲに関連する部分を中心に求めることができた．この結果から，これらの領域では，完全同期で血圧が変化するよりは，時間ずれを伴って変化する場合が多いことを示している．血圧推移や塩分，野菜の変化と比較すると，とくに YG→GY，YG→YR，YR→YR の推移が顕著に飛び出ていることが分かった．より定量的に推移をみるため，c-ratio を使って特徴的な推移を示す部分を強調させた．結果を図６に示す．

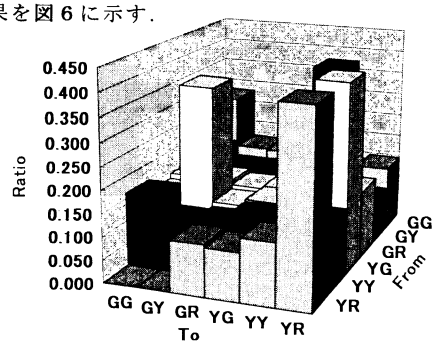

図６：同期・準同期結合パターンの推移頻度比

7.3. 得られた特徴的パターン

完全同期の制約パターンを導入した系列パターンマイニングにより求まったパターンを絞込んだ結果，以下の１つのパターンに絞られた．この再構成前の完全同期パターンでは，血圧がまったく問題ない受診者において，複数の食生活習慣（塩分，野菜）が完全同期してともに悪化すると血圧も悪化することが分かる．

●同期結合パターン：

<s1（c-ratio=1.563, sup=152）>:

塩分 = 無 野菜 = 毎日 血圧 = GG	→	塩分 = 週３・４日 野菜 = 週３・４日 血圧 = GR

次に，6 章のパターン再構成手順に基づいて血圧・塩分の頻出系列パターンと，野菜の頻出系列パターンとを依存関係によって結合させた結果，４つの準同期パターンを求めることができた．

●準同期結合パターン：

<qs1（c-ratio=3.865, sup=181）>:

血圧 = YG 塩分 = 無	→	血圧 = YR 塩分 = 週3･4日
野菜 = 毎日		野菜 = 週3･4日

<qs2（c-ratio=1.563, sup=152）>:

血圧 = GG 塩分 = 無	→	血圧 = GR 塩分 = 週3･4日
野菜 = 毎日		野菜 = 週3･4日

<qs3（c-ratio=3.641, sup=175）>:

血圧 = GY 塩分 = 無	→	血圧 = YY 塩分 = 週3･4日
野菜 = 毎日		野菜 = 週3･4日

<qs4（c-ratio=0.355, sup=326）>:

血圧 = YG 塩分 = 無	→	血圧 = GY 塩分 = 週3･4日
野菜 = 毎日		野菜 = 週3･4日

8. 考察

完全に同期して複数の生活習慣が同時に変化したときの血圧変化について，系列パターンマイニングを用いて解析したところ，同期結合パターンに示すように非常に限られたパターンのみが生成された．これらの推移パターンの頻度分布のパターンは血圧単独の頻度分布のパターンとほとんど変わらないものであった．しかしながら，パターン結合による再構成を用いて求めた準同期結合パターンを求めたところ，血圧単独の頻度分布とは異なる非対称な推移パターンを見つけることができた（図５）．

現在は推移頻度マトリクスの対称性に着目して特長的パターンを算出している．しかしながら，悪化・改善は単に対称ではなく，ヒステリシスを持っている可能性もある．つまり，一旦悪化すると改善しにくくなるといった可能性も考えられる．本研究では悪化・改善をみるために対称性を仮定したが，今後はこのような性質を考慮する必要があると考えられる．

9. 今後の展開

本研究における，依存関係を用いた系列パターン再構成による特徴的な時系列パターンの発見方法は，定期健康診断データなどの年度単位での時系列データのみならず，多変量・多次元センサデータからの特徴的なパターンの発見にも応用可能であると考えられる．今後はセンサデータからの特徴的パターンの発見についても考察し，本方法の有効性を検証する予定である．

文　献

[1] Agrawal, R., and Srikant, R., “Mining Sequential Patterns”, In Proceedings of the 11th International Conference on Data Engineering, IEEE, pp.3-17, 1996.

[2] 植野 研，古川 康一，“ピークタイミングシナジーによる動作スキル理解”，人工知能学会論文誌，Vol.20, No.3, pp.237-246, 2005.

[3] 植野 研，櫻井 茂明，折原 良平，“時間間隔を考慮した営業日報からの系列パターン抽出”，第 3 回情報科学技術フォーラム論文集 (FIT2004)，pp.339-340, 2004.

[4] Sakurai, S., and Ueno, K., “Analysis of Daily Business Reports Based on Sequential Text Mining Method”, In Proceedings of the International Conference on Systems, Man and Cybernetics (IEEE SMC'04), 2004.

[5] Srikant, R., and Agrawal, R., “Mining Sequential Patterns: Generalizations and Performance Improvements”, In Proceedings of the 5th International Conference on Extending Databse Technology (EDBT'96), pp.3-17, 1996.

[6] 須賀 万智，杉森 裕樹，飯田 行恭，吉田 勝美，“職域の定期健診データによる中高年男性の高血圧発症にかかわる要因の解析”，日本公衛誌，pp.543-549, Vol.48, No.7, 2001.

[7] 鈴木 英之進，渡辺 健志，山田 悠，十見 昌俊，大島 宗哲，鍾 寧，横井 英人，高林 克日己，“例外性発見に基づくスパイラル的アクティブマイニング”，人工知能学会誌，Vol. 20, No. 2, pp. 188-195, 2005.