C2-8

時系列データによる時制クラスの発見

本吉 正博[1] 渡辺 浩平[1]
三浦 孝夫[1] 塩谷 勇[2]

[1] 法政大学工学部電気電子工学科, 東京都小金井市梶野町 3-7-2,
E-mail: c9843153@k.hosei.ac.jp
[2] 産能大学経営情報学部, 神奈川県伊勢原市上粕屋 1563,
E-mail: shioya@mi.sanno.ac.jp

Abstract. 本稿の目的は, 時系列データをモデル化する**時制クラススキーマ**を発見することである. この問題は, 時制データマイニングの視点から考えると, 時系列データの**離散化**や集約, またはノイズ除去の問題に置き換えられる. 時制データベースの視点から考えると, どのようにデータ表現をするか, どのように時制スキーマとして潜在的な意味を獲得するかが論点である. 時間軸を伴った時制オブジェクトの集合（これを**ログ**と呼ぶ）を調べ, データを記述するための時制頻出クラスの概念を説明する. 本稿の主な成果として, 時制データの時区間への分割と, それにより得られた時制頻出クラスの一意性と存在を論理的に示せた. 実際の時系列データに適用できることを実験を行い, 検証する.
キーワード: データマイニング, スキーマ発見, 時制スキーマ, 時系列データ

1 時系列データのモデル化

時制データの扱いは, データマイニングの主要な論点のひとつである. 本稿では, データベースを用いることによって, 時系列データをモデル化する方法を述べる. 時系列データは, 工学（気象情報, 遺伝学）, 医学（患者の体温）から金融（製品販売, 株式）まで様々な応用分野で分析されている. 本稿ではこれらの中で主にトランザクションを意識した, 時系列データを用いる。ある店舗の販売記録の場合, 商品名と売上時刻のリストのようなものである. 時系列データは時間の流れにより変化するので, その変化の動向を調べることは, データを分析する上で極めて重要な要素である。

データベース設計の視点から見ると, 時系列データを記述するためのスキーマ（**メタオブジェクト**）を得ることが分析の目的である. 言い換えれば, **データベース設計**とは, 離散スキーマに関して変化の動向を記述することである.

どのようにすれば時系列データを集約できるだろうか. もし, 固定長の時間帯（区間）を決めれば, 細かい並びにデータを分解することができる。しかし大域的動向と局所的動向との両方を捕らえることは難しい. 荒い区間の場合, 局所的動向を見落とすであろうし, 一方細かい区間の場合は大域的動向を見落とすだろう. ノイズを受ける事もある. さらに大事な事は, 前もってイベントが予測できないと, 適当な大きさの区間を決められないということである.

知識抽出を行なうには, ログを用いて時制オブジェクトの頻度を数え, 頻出でない（めったに出現しない）オブジェクトを不要と考え, 取り去る. このステップは頻度の低いオブジェクトをノイズと仮定した, **ノイズ除去**ともいえる. その反面, ある区間でだけ大量に発生するオブジェクトを見落とす可能性もある[3] .

[3] 本稿は頻度に基づいた理論だが, 頻度の代わりに**コスト**に基づいて論議を発展させることもできる.

本稿では, 時制頻出クラスの獲得と分割について述べる. 時制頻出クラスの基本的な性質を示し, **適正**な分割について述べる. そして, 適正な時制頻出クラス列が一意に存在し、どんな時制頻出クラスでも一意に得られることを述べる.

次の章では, ログを用いて時制の知識を**時制頻出クラス**として抽出する方法について述べる. 第 3 章と第 4 章では, 時制頻出クラスの獲得と分割について述べる. 第 5 章では, 実験とその結果を示した後, 第 6 章で本稿をまとめる.

2 ログからの知識抽出

ログから 潜在的な情報を抽出するために, 時制オブジェクトの頻度を調べて, 頻度の低いオブジェクトを捨てる. 非時制オブジェクトの場合の方法を簡単に述べる.

ログ L の要素は、オブジェクトの集合 O の部分多重集合である。集合 LZ を次のように定義する.

$$LZ = \{\langle e, n\rangle | e \in L, L \text{ に } e \text{ が } n \text{ 回出現 }\}$$

LZ は L 上で一意である. $e \notin L$ ならば $\langle e, 0\rangle \in LZ$ とする. **最小サポート** s は実数 $0.0 \leq s \leq 1.0$ の間とする.

もし $\langle e, n\rangle \in LZ$ と $n > s \times |L|$(ただし $|L|$ はLのタプル数), 両方を満たすならば要素 $e \in L$ は**頻出**であるとする.

また,$D \subseteq L$ に関して以下を満たすなら,D は**頻出**クラスとする.

(1) L のなかですべての D の要素が頻出
(2) どの $x, y \in D$ も $x \supseteq y$ ならば,$x = y$[4]
(3) どの $x \in L$ も L で x が頻出ならば,$y \supseteq x$ であるような $y \in D$ が存在する.

(3) の性質を満たす時,D は L を**カバー**するという.

このログ L に一意な頻出クラス D が存在するということは興味深いことである.

区間 T 上の時制オブジェクトの集合 O があるとき, **時制ログ**はOの部分多重集合で,L(T) と表す. 集合 LZ を次のように定義する.

$$LZ = \{\langle e, n\rangle | e \in L(T), time(e) \in T, L \text{ に } e \text{ が } n \text{ 回出現する }\}$$

LZ は L 上で一意であり, もし $e \notin L$ ならば $\langle e, 0\rangle \in LZ$ とする. O をベキ集合,$e \in L$ を集合値とし, 時制クラスに定理 1 を適用する. もし $\langle e, n\rangle \in LZ$ かつ,$n > s \times |L|$ ならば, 要素 e は L(T) において**頻出**であるとする.

$S \subseteq T$ のとき、区間 T,S において時制ログ L(T),L(S) があるとき, 上の**時制部分ログ** $L[S]$ は, 次のように定義された部分多重集合である.

$$L[S] = \{e \in L(T) | time(e) \in S\}$$

L(T) は空ではなく, L が隙間なくに密集しているとする. $L[T] = L(T)$ とする. 定義より,$T = T_1 * .. * T_n$ があるとき,$L[T_i] \cap L[T_j]$ は $i \neq j$ において空であり,$|L(T)| = |L[T_1]| + |L[T_2]| + .. + |L[T_n]|$ である.

$S \subseteq T$ のとき、区間 T,S と時制ログ $L = L(T)$ において T 上の**時制頻出クラス** C(T) を次のように定義する.

$$C(T) = \{e \in L(T) | e \text{ は } L(T) \text{ で頻出 }\}$$

また,$C[S]$ と表される S 上の**時制頻出部分クラス**を次のように定義する.

[4] ここで x と y は集合値であるとする.

$$C[S] = \{e \in L[S] | e \text{ は } L[S] \text{ で頻出 }\} \cap C$$

ここで C は空ではないが $C[S]$ は空でもよい. 定義より $C[S]$ は C の部分集合であり, 故に T 上で頻出な要素だけに注目すればよい. $C'[S]$ を下記のような集合とする.

$$C'[S] = \{e \in L[S] | e \text{ は } L[S] \text{ で頻出 }\}$$

明らかに $C[S] \subseteq C'[S]$ であるが $C[S] = C'[S]$ は次の例にも示すが常には成り立たない. すなわち, 局所的な頻出時制オブジェクトでも全体では頻出でないことがある. これは局所的動向でなく, 大域的動向に注目する理由である.

例 1 $C[S] \neq C'[S]$ である例を示す. 次のような時制ログがあるとする.

	T_1	T_2
a	1	0
b	0	2

時制オブジェクト a は T_1 で 1 度出現し,b は T_2 で 2 度出現し, その他の出現はない. $s = 0.60, T = T_1 * T_2$ とすると $C(T) = \{b\}$. その時,$C'[T_2] = C[T_2] = \{b\}$ であるが $C'[T_1] = \{a\}, C[T_1] = \{\}$ である. □.

3 時制クラスの獲得

この章では, 時制クラスの基本的な性質について述べる.

定理 1 T を区間とし,$T = T_1 * .. * T_n$ の分割を $T_1, .., T_n$ とする. また, C を T 上の時制頻出クラスとする. その時,$C = C[T_1] \cup ... \cup C[T_n]$ である.

(証明) C ⊇ C[Ti] なので,$C \supseteq C[T_1] \cup ... \cup C[T_n]$ は自明. 逆にどの C[Ti] にも含まれない e は C に含まれないことを示す. e ∈ C[Ti] でないというのは,e ∈ C でないかまたは L[Ti] で頻出でないということである. 前者なら終わり. e ∈ C かつ L[Ti] で頻出でない場合を考えると定義より $\langle e, m_i \rangle \in LZ[T_i]$ でかつ $m_i \leq s \times |L[T_i]|$ これがどの i でも成り立ち,$|L(T)| = |L[T1]| + .. + |L[Tn]|$ だから $m_1 + .. + m_n \leq s \times (|L[T1]| + .. + |L[Tn]|) = s \times |L|$. $\langle e, m \rangle \in LZ$ に対して, $\langle e, m \rangle = \langle e, m1 \rangle + .. + \langle e, mn \rangle$ であるから $m > s \times |L|$ でない. つまり $e \in C$ ではない. これは矛盾する. □.

この定理は,C が T の分割に伴って分割が可能であることを意味している. 例えば,$T = T_1 * .. * T_n, T_1 = S_1 * S_2$ のとき $C(T) = (C[S_1] \cup C[S_2]) \cup ... \cup C[T_n]$ が成り立つ.

定理 2 $T = T_1 * .. * T_n$ の分割を $T_1, .., T_n$ とする. また,C を T 上の時制頻出クラスとする. $1 \leq i < n$ とする. このとき,
(a) $C[T_i] \cap C[T_{i+1}] \subseteq C[T_i * T_{i+1}]$
(b) $C[T_i * T_{i+1}] \subseteq C[T_i] \cup C[T_{i+1}]$

(証明) $i = 1$ とする.

(a) e ∈ C[T1] ∩ C[T2] とする. このとき,$\langle e, m_1 \rangle \in LZ[T_1], \langle e, m_2 \rangle \in LZ[T_2]$, $m_1 > s \times |L[T_1]|, m_2 > s \times |L[T_2]|$ である m_1, m_2 が存在する. $\langle e, m \rangle \in LZ[T_1 * T_2]$ とすれば,$m = m_1 + m_2 > s \times (|L[T_1]| + |L[T_2]|) = s \times |L[T_1 * T_2]|$ なので,e は $L[T_1 * T_2]$ で頻出であるから,$e \in C[T_1 * T_2]$.

(b) 性質 (a) と同様に考えると,$e \in C[T_1], e \in C[T_2]$ が共に成立しないとき $e \notin C[T_1 * T_2]$ である. □

上記の定理には, 次の例で示す s ように, 一般に等号は成り立たない.

例 2 $s = 0.33$ とし, 次のようなログ L がある：

	T_1	T_2	T_3
a	1	0	2
b	0	1	2
c	1	1	1

C は $\{a, b, c\}$ である. $C[T_1] = \{a, c\}$, $C[T_2] = \{b, c\}$, $C[T_3] = \{a, b\}$, $C[T_1 * T_2] = \{c\}$, $C[T_{23}] = \{b\}$ である. 特に $C[T_{12}] \neq C[T_1] \cup C[T_2]$ を意味する C[T1*T2] = c, C[T1] ∪ C[T2]=a,b,c. また,$C[T_1] \cap C[T_{23}] = \{a, c\} \cap \{b\} = \phi$, $C[T_1 * T_{23}] = C(T) = \{a, b, c\}$ であるので $C[T_1] \cap C[T_{23}] \neq C[T_1 * T_{23}]$ である. □

区間を拡張するために隣接したクラスを結合することができる.

定理 3 $T = T_1 * .. * T_n$ の分割を $T_1, .., T_n$ とする. また, C を T 上の時制頻出クラスとする. その時,$C[T_i] = C[T_{i+1}]$ なら $C[T_i] = C[T_i * T_{i+1}]$ である.

(証明) 定理 2 より,$C[T_i] \cup C[T_{i+1}] \supseteq C[T_i * T_{i+1}] \supseteq C[T_i] \cap C[T_{i+1}]$. $C[T_i] = C[T_{i+1}]$ なので,$C[T_i] = C[T_i] \cup C[T_{i+1}] = C[T_i] \cap C[T_{i+1}]$ であり定理は成り立つ. □

追加説明をする. 逆の例があり,$C[T_1] = C[T_1 * T_2] \neq C[T_2]$ なのである. これについては後述する. つまり, 連続する時区間で変化が無ければこの隣接区間を合併しても本質的にかわらない。しかし、その逆は成り立たない.

4 時制クラスの分割

前章では, 時制頻出クラスを**記述**する時制頻出部分クラス列が 1 つ以上存在することを示した. この章では, **記述**の定義とクラス列が満たす性質について述べる.

初めに, 記述の定義について述べる. $T = T_1 * .. * T_n$ とする. L を T 上のログ, C を T 上の時制頻出クラス,C_i を T_i 上の時制クラスとする. もしそれらが, 下記をすべて満足するなら,$C1, .., C_n$ は C を記述すると定義する.

(1) どの C_i も T_i 上の時制頻出部分クラス
(2) $C = C_1 \cup ... \cup C_n$ ただし $n > 1$
(3) $C_i \neq C_{i+1}$, $i = 1, .., n-1$

ただし,C_i は空でもかまわない.

定理 4 任意の C(T), $T = T_1 * .. * T_n$ に対して,C を記述する $T_1, .., T_n$ 上の $C1, .., C_n$ が存在する

(証明) $C_i = C[T_i]$ とする. 明らかに $C[T_i]$ は（1）（2）を満たす. $C[T_i] = C[T_{i+1}]$ となるような i があれば定理 4 より $C[T_i]$ は $C[T_{i+1}]$ と等価であるので,Ti,T(i+1) を合併できる. □

C を記述する時制頻出クラス列は次に示すようにいくつかある.

例 3 性質を調べるためにいくつかの例を示す.

	T_1	T_2	T_3
a	1	0	1
b	0	2	0

(a)

	T_1	T_2	T_3	T_4
a	1	0	0	1
b	0	1	1	0

(b)

	T_1	T_2	T_3	T_4
a	1	1	0	0
b	0	0	1	1

(c)

（1）最小サポートを 0.50. ログを上記の表（a）とする.$C = \{a, b\}$, $C[T_1] = \{a\}$, $C[T_2] = \{b\}$, $C[T_3] = \{a\}$,$C[T_1 * T_2] = \{b\}$,$C[T_2 * T_3] = \{b\}$ である. その時, $\{C[T_{12}], C[T_3]\}$, $\{C[T_1], C[T_{23}]\}$,$\{C[T_1], C[T_2], C[T_3]\}$ は C を記述する. そして, 最も状況を忠実に反映しているのは, $\{C[T_1], C[T_2], C[T_3]\}$ である.

（2）最小サポートを 0.50. ログを上記の表（b）とする.$C = \{a, b\}$, $C[T_1] = \{a\}$, $C[T_2] = C[T_3] = \{b\}$, $C[T_4] = \{a\}$, $C[T_1 * T_2] = \{a, b\}$, $C[T_2 * T_3] = b$, $C[T_3 * T_4] = \{a, b\}$, $C[T_1 * T_2 * T_3] = \{b\}$, $C[T_2 * T_3 * T_4] = \{b\}$ である. その時, $\{C_1, C_{234}\}$,$\{C[T_{123}], C[T_4]\}$, $\{C[T_{12}], C[T_3], C[T_4]\}$, $\{C[T_1], C[T_{23}], C[T_4]\}$, $\{C[T_1], C[T_2], C[T_{34}]\}$ はCを記述する. しかし $\{C[T_{12}], C[T_{34}]\}$ は $C[T_{12}] = C[T_{34}]$ なので定義（3）を満たさず, よって記述しない.$\{C[T_1], C[T_2], C[T_3], C[T_4]\}$ は $C[T_2] = C[T_3]$ なので, 同様に記述しない. 最も状況を忠実に反映しているのは, $\{C[T_1], C[T_{23}], C[T_4]\}$ である.

（3）最小サポートを 0.50. ログを上記の表（c）とする.$C = \{a, b\}$,$C[T_1] = C[T_2] = \{a\}$, $C[T_3] = C[T_4] = \{b\}$,$C[T_1 * T_2] = \{a\}$, $C[T_2 * T_3] = \{a, b\}$, $C[T_3 * T_4] = \{b\}$,$C[T_1 * T_2 * T_3] = \{a\}$, $C[T_2 * T_3 * T_4] = \{b\}$ である. その時, $\{C[T_1], C[T_{234}]\}$, $\{C[T_{123}], C[T_4]\}$, $\{C[T_{12}], C[T_{34}]\}$, $\{C[T_1], C[T_{23}], C[T_4]\}$, $\{C[T_1], C[T_2], C[T_{34}]\}$ はC を記述するが,$\{C[T_{12}], C[T_3], C[T_4]\}$,$\{C[T_1], C[T_2], C[T_{34}]\}$,$\{C[T_1], C[T_2], C[T_3], C[T_4]\}$ は記述しない. 最も状況を忠実に反映しているのは, $\{C[T_{12}], C[T_{34}]\}$ である. □

次に, $T_i = S_1 * ... * S_k$ のような T_i の任意の分割 $S_1, .., S_k$ に対して,$j = 1, .., k-1$ で $C_i[S_j] = C_i[S_{j+1}]$ ならば T_i 上の C_i は, Cに関して**適正**であると定義する. また,$i = 1, .., n$ において C_i がCに関して適正ならばクラス列 $C_1, .., C_n$ は**適正**であると定義する.

例 4 適正, 不適正な時制頻出クラス列の例を示す.
(1) 例3（1）で,$\{C_1, C_2, C_3\}$ はCを記述する適正なクラス列である. 同様に例3（2）の $\{C_1, C_{23}, C_4\}$, 例3（3）の $\{C_{12}, C_{34}\}$ は適正である. しかし, その他は適正ではない. 例えば例3（3）の $\{C_1, C_{23}, C_4\}$ はCを記述するが C_{23} が適正でないために, 適正なクラス列ではない.

	T_1	T_2	T_3
a	2	1	1
b	1	1	2
c	0	1	0

(a)

	T_1	T_2	T_3
a	1	1	2
b	1	0	0
c	0	1	0
d	0	0	1

(b)

(2) ログとして表 (a) があり,$s = 0.40$ とする. この時,$C[T_1] = \{a\}, C[T_2] = \{\}, C[T_3] = \{b\}, C[T_{12}] = \{a\}$, $C[T_{23}] = \{b\}$ である. $C[T_{12}]$ は $C[T_1] \cup C[T_2]$ へ,$C[T_{23}]$ は $C[T_2] \cup C[T_3]$ へ分解されるので $\{C\}, \{C[T_{12}], C[T_3]\}, \{C[T_1], C[T_{23}]\}$ のどれも適正とはいえない.

また, ログとして表 (b) があり,$s = 0.51$ とする. $C = C[T_{123}] = \{a\}$, $C[T_1] = C[T_2] = \{\}$, $C[T_3] = \{a\}$, $C[T_{12}] = \{\}$,$C[T_{23}] = \{a\}$ である. $C[T_{12}] = C[T_1] \cup C[T_2]$,$C[T_{23}] = C[T_2] \cup C[T_3]$ と分解でき,$\{C\}$, $\{C[T_1], C[T_{23}]\}$ は適正ではないが,$\{C[T_{12}], C[T_3]\}$ は適正である. □

定理 5 T,L,C をそれぞれ, 区間, ログ, T上の時制頻出クラスとする. この時,$T = T_1 * .. * T_n$ の分割 $T_1, .., T_n$ と, Cを記述する $T_1, .., T_n$ 上の適正な時制頻出クラス列 $C_1, .., C_n$ が存在する.

(証明) T を最小区間 $T_1, .., T_n$ に分割する. $C[T_i] = C[T_{i+1}]$ である限り次々に T_i と T_{i+1} を合併していく. $n < \infty$ なので, これ以上合併できないという連続区間 $S_1, .., S_m$ において時制頻出クラス列 $C_1, .., C_m$ を得る. これが適正であることを示す.

S_i が最小区間なら定義より分解できないので適正. 次に,$S_i = T_a * .. * T_b$ とする. U_1, U_2 を最後に合併する区間とすると $S_i = U_1 * U_2$ であり, 定義より $C[S_i] = C[U_1] = C[U_2]$ である. 帰納法を使えば $C[T_a] = ... = C[T_b] = C[S_i]$ であるから, これは適正条件を満たす. □

定理 6 定理５で T の分割 $T_1, .., T_n$ とＣを記述する適正な時制頻出クラス列 $C_1, .., C_n$ は一意である.

(証明) $T_1, .., T_n$,$S_1, .., S_m$ が共に適正であるとする. $T_1 = S_1$ と $n = m$ を示し, ２つは同一であることを帰納法により証明する.

まず,$T_1 = [t_1..t_2), S_1 = [s_1..s_2)$ とすれば,$T_1 = S_1$ であることを示す. ２つは共に第１区間であるから $t_1 = s_1$ である. そして $T_1 \subseteq S_1$ または $S_1 \subseteq T_1$ である. $T_1 \subseteq S_1$ の場合,$t_2 \leq s_2$ である. $t_2 < s_2$ とすれば,$S_1 = T_1 * U_1$ と分解できる. ここで $U_1 = [t_2..s_2)$ である. 適正な列であることから $C[T_1] = C[U_1] = C[S_1]$ である. $U_1 \cap T_2 \neq \phi$ なので,$C[U_1] = C[T_2]$ となり $C[T_1] \neq C[T_2]$ とは矛盾する. 故に $t_2 = s_2$ である.

次に $n = m$ を示す.n ≠ m とすると,$T_i \neq S_i$ となる最小の i が存在する. $T_i = [t_i..a), S_i = [s_i..b)$ とすれば i が最小なので $t_i = s_i$ である. ここで,$T_i \subseteq S_i$,$a < b$ とすると, 上と同様に矛盾する. □

一意な区間分割と適正な時制頻出クラス列を得るために, 上記の定理を満たすアルゴリズムを次に示す.

1. ログＬの区間をＴとし, wを最小区間とする.
2. wでＴを $T_1, .., T_n$ へ分割する.
3. $C(T)$,$C[T_i]$(ただし $i = 1, .., n$) を計算する.
4. $C[T_i] = C[T_{i+1}]$ である限り次々に T_i と T_{i+1} を合併する.

5 実験

本稿の理論に基づいて行なった実験とその結果を示す. 学生のための就職情報ページの**ウェブログ**の中から１日分を用いる. ログはＩＰアドレスによって分類され, タイムスタンプの順に記録されている. 下にそのなかの一部を示す.

```
......
[13:14:14] /job/ /job/mokuji.htm /job/info.htm /job/BK253.GIF
[13:14:16] /job/2000/M20.htm
[13:14:20] /job/ /job/ /job/mokuji.htm /job/info.htm /job/BK253.GIF ...
[13:14:24] /job/suisen/index.htm
[13:14:25] /job /job/ /job/mokuji.htm /job/info.htm /job/BK253.GIF ...
[13:14:28] /job/suisen/index.htm
[13:14:29] /job/suisen/april25.htm
[13:14:30] /job/suisen/april25.htm
[13:14:37] /job/ /job/ /job/mokuji.htm /job/info.htm /job/BK253.GIF ...
[13:14:45] /job/ /job/mokuji.htm /job/info.htm /job/BK253.GIF ...
.....
```

ユーザーの大部分が学生であり, 日付, 週という項目は無視されるため, 前もってパターンを予想することはできない. サポートは１０%, ２０%, ３０%としてそれぞれ本稿のアルゴリズムを実行し, 時制頻出クラス `C[0:00-24:00]` を調べた.

分類された状態を図１に示す. 白い部分が合併されなかった区間で, 色つきの部分が合併された区間である. 分かりやすいように, 色を複数用いた. ３本あるのは, 最小区間が上から３０分, ６０分, １２０分の場合である.

サポートが１０%, 最小区間が３０分の場合, ２８個の時区間がアルゴリズムによって得られた. この結果は, 最小区間である４８個の区間が, ２８個の区間に分類されたことを示している. 最小区間が６０分の場合, １３個の区間を得た. これは, ２４個の区間が１３個に分類されたことを意味している. １２０分の場合, １２個の区間が６個に分類された. これによると学生が昼食と夕食と（多分）入浴時間を除いた日中（午後）と夜の時間帯が大きく合併されており, なんらかの潜在的な意味がある強いつながりを持った時間帯だということが分かる. 日中と夜の時間帯である事からこの区間は学生が活発に活動する時間帯であることを示していると考えられる.

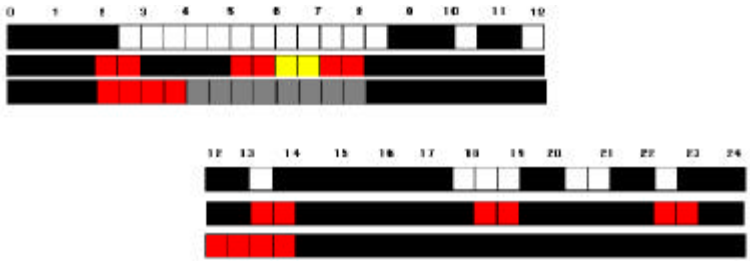

図. 1. サポート 10% (30 分, 60 分,120 分)

サポートを２０%とした場合も, だいたい類似の結果を得た. 最小区間を３０分として４８個から２３個の区間を, ６０分として２４個から１０個の区間を, １２０分として１２個から５個の区間をそれぞれ得た. 分類された状態を図２に示す. 結果は図１と類似した.

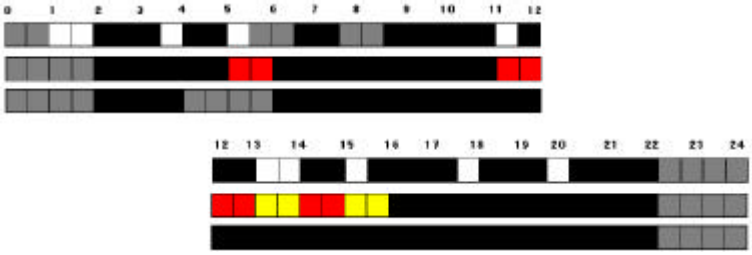

図. 2. サポート 20% (30 分, 60 分 ,120 分)

他のケースと比較して, ３０%の場合はまったく異なった結果を示した. 最小区間を３０分として４８個から５個の区間を, ６０分として２４個から３個の区間を, １２０分として１２個からはたった１個の区間を得た. これらの結果を図３に示す.

たかだか早朝と昼食の時間だけを抽出したが, 特徴を抽出できる十分なアイテムは得られなかった.

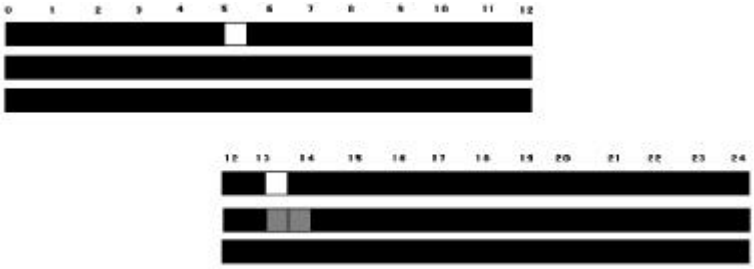

図. 3. サポート 30% (30 分, 60 分,120 分)

WebLog を用いた実験を要約する. 区間は４８, ２４, １２個より合併された. その時の減少した区間数の比率を表に示す. 本稿のアルゴリズムによって区間が分割され, 適正な頻出部分クラス列が得られた. もちろん, より低いサポートを与えれば, より小さな時制頻出部分クラスが生成されると推測できる. ３０%の場合が他とまったく異なっているが, １０%と２０%のサポートは類似の減少を示している. サポートが高いほど区間はより減少する. 最小区間がより大きいと区間の減少が下げるがたいした差はなかった.

support	30m (48)	60m (24)	120m (12)
10%	28(58%)	13(54%)	6(50%)
20%	23(48%)	10(42%)	5(42%)
30%	5(10%)	3(13%)	1(8%)

6 結論

本稿で, 時系列データのモデル化と, **時制クラススキーマ**の発見について述べた.

時間軸を伴った時制オブジェクトのデータを調べる事によって, それらを記述する時制頻出クラス列を得ることができた. また, 区間分割とそれによる時制クラス列の一意性と存在性を論理的に示すことが出来た. 本稿の理論を検証するいくつかの実験結果を示した.

我々はすでに確率論を基盤にエルゴード理論の技術について研究しており, それとの統合法について研究している [2].

References

1. Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, VLDB94, pp.487-499, 1994
2. Miura,T. and Shioya, I. et al.: Behavior Discovery as Database Scheme Design, *TIME*, pp.115-122, 2000

This article was processed using the LaTeX macro package with LLNCS style