時系列データからの時制クラスの発見

本吉正博[†] 三浦孝夫[†] 塩谷勇[††]

本研究では離散型時系列データ集合からの時制クラスの発見方法を論じる．時制データマイニングの観点からは，この問題は時系列データの離散化あるいは集約化と考えることができる．またノイズ除去問題であるとも見なせる．他方，時制データベースの観点からいえば，この問題は，どのようにしてデータを表現しどのようにして潜在的な意味をスキーマとして獲得するか，いい換えるとデータベーススキーマ発見問題ととらえることができる．時間軸に沿って時制オブジェクト集合（ログ）が与えられたとき，これらを解析し記述するため，‘時制頻出クラス’概念を導入する．本研究の主な結果として，時区間の分割と関連した時制クラスがつねにただ１つ存在することを示すことができる．また，実際の時系列データを用いて実験を行い，本アプローチの有効性を示す．

Mining Temporal Classes from Time Series Data

Masahiro Motoyoshi,[†] Takao Miura[†] and Isamu Shioya[††]

In this investigation, we discuss how to mine *Temporal Class Schemes* to model a collection of time series data. From the viewpoint of temporal data mining, this problem can be seen as *discretizing* time series data or aggregating them. Also this can be considered as screening (or noise filtering). From the viewpoint of temporal databases, the issue is how we represent the data and how we can obtain intensional aspects as temporal schemes. In other words, we discuss scheme discovery for temporal data. Given a collection of temporal objects along with time axis (called *log*), we examine the data and we introduce a notion of temporal frequent classes to describe them. As the main results of this investigation, we can show that there exists one and only one interval decomposition and the temporal classes related to them. Also we give experimental results that prove the feasibility to time series data.

1. 時系列データのモデル化

データマイニングでの主要な論点の１つは，時制情報を含むデータをどのように取り扱うかにある[3)]．本稿では，データベースを利用して離散型時系列データをモデル化する方法を論じる．時系列データは工学分野（たとえば，センサを利用した監視業務，遺伝学），医療分野（患者の体温とその効果），金融分野（製品販売，株式売買）等で幅広い応用分野に存在し利用されている．その分析能力は，高度な意思決定を行うためにはきわめて重要である．本研究では，離散型時系列データは主にトランザクションに基づくものとする．たとえば，顧客が購入した項目のリストが，時間情報とともに分析対象になることを想定すればよい．このデータは時間依存する可能性があり，時間軸に沿ってどのように変化するかを調べたい．時間は全順序を有しており，すべての対象値は時間順に整列されている．時系列データの分析は，なぜ，どのような環境でそのようなデータを得たかという点に注目することが多い．このため，離散的かつ静的な潜在的意味を抽出することが主要な興味となり，細かな時間情報は無視される．

時制データベース設計の観点からいえば，この分析はスキーマの獲得を目標にしている．つまり，時系列データを検査し，離散化されたデータの特性について，変化の潜在的意味を記述することが目的である．

本研究では，オブジェクトが時刻印を有するとき，これを時制オブジェクトと呼ぶ☆．時制データベースには，時間的な意味を有するオブジェクト集合が含まれる．たとえば，‘ランチ’データベースにはヨーグルトは含まれないが，サンドウィッチ，ピザや点心は含んでいる．そのデータベーススキーマではある種の時

† 法政大学工学研究科電気工学専攻
Department of Electronics, Electrical & Computer Engineering, Hosei University
†† 産能大学経営情報学部
Department of Management and Informatics, SANNO University

☆ 一般に並びを有するオブジェクトを時制列，特に連続で実数値の並びを時系列という[3)]．本稿では単純に時刻印を有するオブジェクトと考える．

制に関する制約をともなっており，オブジェクトはその制約を充足する必要がある．

時制オブジェクトの頻度を数えることにより，高頻度（頻出）でないオブジェクトを不要と見なし取り去ることを考える．これは低頻度のオブジェクトを雑音と見なし，除去することに相当する．むろん，この方法では，ある時区間でだけ集中して生じたオブジェクトを特徴づけることができない．本研究では，クラスに時区間を対応させた‘時制クラス’を論じる．基本的で重要な性質を示し，‘適正な時区間分割’の概念を導入する．本研究の主要な結果は，与えられた時制頻出クラスに関し，適正な時区間分割がただ 1 つ存在するということである．したがって時制頻出クラスを最適な形で得ることができる．

離散型の時制データに関しては，高次元連続データ・ストリームデータの直接のモデル化が困難であることはよく知られている[7]．過去に時制データベースおよび時制データマイニングの分野で研究が試みられている．前者では時制データモデリング[6),16)]や時制質問処理[8)]，SQL の拡張[13)]，TQuel 言語の提案[15)]，関係代数の拡張[9)]等が代表的である．一方，時制データマイニング分野における考察は連続型時制データに比して多くない[3)]．時制データの類似度の定義や同時関係の発見について時制を扱うものはあまり知られていない．これに対して，時制データのクラス分類については，確率論的なアプローチが可能である．たとえばストリームデータに関する決定木生成の研究では，サンプリングアプローチ[4)]やマルコフ過程の下でのエルゴード性分析によるアプローチ[12)]等が知られる．本研究では，離散型データに対して時制クラス発見を行うため，データマイニング手法を適用する．この意味で，境界領域的な解法を提案するものである．

次章では，本研究で用いる時制データベースの定義をする．3 章では，ログから時制頻出クラスを抽出する方法を論じる．4 章と 5 章では，時制頻出クラスの分析とその性質を示す．6 章ではいくつかの実験結果を示す．最後に本研究の要約を述べる．

2. 時制データの記述

‘オブジェクト’とは対象世界に存在する‘もの’（*thing*）である．‘関連’は個々のオブジェクト間の結びつきの表現 $\langle e_1,..,e_n\rangle$ であり，構成するオブジェクトとは別個に存在する．‘オブジェクトクラス’により，共通特性を有するオブジェクト集合を記述するための概念を表す．各クラスに対して属性集合が対応し，これによりオブジェクトを記述することができる．‘述語’は関連を表現するための潜在的な意味を示し，クラスと同様に共通の属性集合が対応する．

‘時制データモデル’はクラス，述語，オブジェクトおよび関連から成り，また一貫性を保持するための時制制約を含む[5)]．オブジェクトは時制的であってもなくてもよい．時制オブジェクト e とは生成時刻印 $time(e)$ を含むオブジェクトである．同様に時制関連 $\langle e_1,..,e_n\rangle$ も時制情報を有し，その時刻印を $time(\langle e_1..e_n\rangle)$ で表す．時制関連の構成オブジェクト e_i は時制的でなくてもよい．以下では O，R をそれぞれ時制オブジェクト，時制関連の集合とする．

重要な概念に‘時区間’（interval）がある[2)]．時区間とは，集合 $[t_1..t_2)$，すなわち $t_1 < t_2 < \infty$ であり，$t \in [t_1..t_2) \Leftrightarrow t_1 \leq t < t_2$ で定義される半閉時区間である．2 つの時区間 T= $[t_1..t_2)$, S= $[s_1..s_2)$ に対し，$t_1 < t_2 \leq s_1 < s_2$ ならば‘T は S より先行する’といい，T < S と表す．同様に，$t_1 \leq s_1 < s_2 \leq t_2$ ならば‘T は S を含む’といい，T $\supseteq$ S あるいは S $\subseteq T$ と表す．$t_1 \leq s_1 < t_2$ または $s_1 \leq t_1 < s_2$ の少なくとも一方が成り立つとき，‘T は S と交差する’といい，$t_2 = s_1$ ならば‘S は T に続く’という．S が T に続くとき合併した時区間を $T * S = [t_1..s_2)$ と定義する．同様に $\mathrm{T} = T_1 * .. * T_n$ を定義することができる．逆に $\mathrm{T} = T_1 * .. * T_n$ ならば $T_1,..,T_n$ を T の‘分割’という．混乱がない限り T の分割 $T_1,..,T_n$ を $T_1 * .. * T_n$ とも表し，断らない限り添え字と分割を対応させる．たとえば T_{12} によって $T_1 * T_2$ を，T_{123} によって $T_1 * T_2 * T_3$ を表す．T に 2 つの分割 $\mathcal{T} = \{T_1,..,T_n\}$，$\mathcal{S} = \{S_1,..,S_m\}$ があり，どの T_i もいくつかの S_j の合併で記述できるなら，$\mathcal{S}$ は $\mathcal{T}$ より‘詳細である’という．

時制クラス C とは，時区間 T を属性として有し，C のオブジェクト e は $time(e) \in \mathrm{T}$ を満足する．$C(T)$ とも表す．時制述語 P についても同様であり，$P(T)$ とも表す．関連 $\langle e_1,..,e_n\rangle$ についても同様に $time(\langle e_1,..,e_n\rangle) \in T$ を満たす必要がある．以下ではこの時制制約に基づいた時制クラス・述語の設計・発見手法を論じる．

時制スキーマを得るためには，（時制）クラスや（時制）述語だけではなく，時区間を得なければならない．しかし，与えられたデータに対しスキーマはどのような性質を有するべきかという問いに答えることは容易ではない．本稿では，時制スキーマは正確で簡潔な様相を反映するとの立場から，スキーマを得るには，主たる影響を有さないと判断されるデータ（雑音データ）を考察から除外する．むろん，この判断は，クラスに

対する事象や構造に関して十分に合理的であると判断されなければならない．

3. ログからの知識抽出

ログから潜在的な意味情報を抽出するために，時制オブジェクトの頻度を調べ低頻度のオブジェクトを考察から除外する．まずオブジェクトの発生頻度に基づくデータマイニング技法について要約する．

オブジェクト集合 O が与えられたとき，O の部分集合を要素とする多重集合 L を考えログと呼ぶ．L 上の集合 LZ を次のように定義する．

$LZ = \{\langle e, n\rangle \mid e \in O, L \text{ に } e \text{ が } n \text{ 回出現する }\}$

すなわち LZ は L の 1 項目集合で，各要素に出現数を対応させたものである[1)]．L に対して LZ は一意であることに注意したい．最小サポート s とは実数 $0.0 \leq s \leq 1.0$ とする．

要素 $e \in O$ に対して e が L で頻出であるとは，$\langle e, n\rangle \in LZ$ かつ $n > s \times |L|$ が成り立つときをいう．ここで $|L|$ は多重集合 L の濃度を表す．L で頻出なすべての要素の集合 D を ‘頻出クラス’ と呼ぶ．

時区間 T 上の時制オブジェクト集合 O に対し，時制ログ L(T)（または L）とは，O の部分集合を要素とする多重集合である．L(T) に対して次のように集合 LZ(T) を定義する．

$\{\langle e, n\rangle \mid e \in O, L \text{ に } e \text{ が } n \text{ 回出現 }\}$

例 1 6 つの要素からなるログ L(T) と，それぞれの $time$ 値を示す．ここで $a, b, c \in O$, $t1, .., t6$ は時区間 T に含まれる．右には L に対する LZ(T) を示す．

ログ要素	$time$
$\{a, c\}$	$t1$
$\{b\}$	$t2$
$\{c\}$	$t3$
$\{a, b\}$	$t4$
$\{a, b\}$	$t5$
$\{c\}$	$t6$

オブジェクト	出現回数
a	3
b	3
b	3

□

時区間 T，S が $S \subseteq T$ を満たし，時制ログ L(T) が与えられたとき，S 上の部分時制ログ L[S] とは次のように定義される多重集合である．

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

ここで L(T) には多数の要素が含まれ，この結果 L[S] は空集合ではないとする．定義より L[T]=L(T) である．$T = T_1 * .. * T_n$ であるとき，$|L(T)| = |L[T_1]| + |L[T_2]| + .. + |L[T_n]|$ である．

時区間 T，S が S $\subseteq$ T を満たし，時制ログ L(T) が与えられているとする．L(T) で T 上の時制頻出クラス C(T) を次のように定義する．

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

S 上の部分時制頻出クラス C[S] を次のように定義する．

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

C が空集合でなくても C[S] が空となることがある．定義より C[S] は C の部分集合であるので，以下では T 上で頻出な要素だけに注目する．実際，C'[S] を次のような集合とする．

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

明らかに C[S] $\subseteq C'$[S] であるが C[S]=C'[S] は（次に示されるように）つねには成り立たない．すなわち，局所的な頻出時制オブジェクトは全域的に頻出とは限らない．これが局所的な意味を論じないで，全域的な時区間に注目する理由である．

例 2 局所的な頻出時制オブジェクトが全域的に頻出とならない例を示す．T=$T_1 * T_2$ とし，T 上の時制ログが次で与えられているとする：

要素	$time$
$\{a\}$	$t1$
$\{b\}$	$t2$
$\{b\}$	$t3$

ここで $t1$, $t2$, $t3$ は各要素の $time$ 値であり $t1 \in T_1$, $t2 \in T_2$, $t3 \in T_2$ のとき，$LZ(T_1)$, $LZ(T_2)$ をまとめて次のように表す．

	T_1	T_2
a	1	0
b	0	2

時制オブジェクト a は T_1 に 1 度出現し，b は T_2 に 2 度出現しているにすぎない．$s = 0.60$，T において C(T) = $\{b\}$, $C'[T_2] = C[T_2] = \{b\}$ であるが $C'[T_1] = \{a\}$, $C[T_1] = \{\}$ である． □

4. 時制クラスの抽出

この章では時制クラスの基本的な性質を示す．

定理 1 T を時区間とし T 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．このとき $C = C[T_1] \cup ... \cup C[T_n]$ が成り立つ[14)]．

（証明）$i = 1, .., n$ に対して，$C \supseteq C[T_i]$ である．逆を証明するため，どの $C[T_i]$ にも含まれない e は C に含まれないことをいう．$e \in C[T_i]$ でないから，$e \in C$ でないか，あるいは $L[T_i]$ で頻出ではない．前者なら

証明できたので，$e \in C$ かつ $L[T_i]$ で頻出でないとする．

定義より $\langle e, m_i\rangle \in LZ[T_i]$ でかつ $m_i \leq s \times |L[T_i]|$ を得る．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，$\langle e, m\rangle = \langle e, m_1\rangle + .. + \langle e, m_n\rangle$ であるから $m > s \times |L|$ でない．つまり $e \in C$ ではない．これは矛盾である． □.

T 上の時制クラス C において T=$T_1 * .. * T_n$, $T_1 = S_1 * S_2$ のとき C= $(C[S_1] \cup C[S_2]) \cup C[T_2] \cup .. \cup C[T_n]$ が成り立つことが容易に分かる．さらに $T = T_1 * .. * T_n$ という性質を一般化して，分割が排他的なら上記の性質を示すことができる．

定理 2 T を時区間，その分割 $T = T_1 * .. * T_n$，C を T 上の時制頻出クラスとする．このとき，$1 \leq i < n$ に対して
(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）一般性を失わずに $i = 1$ としてよい．
(a) $e \in C[T_1] \cap C[T_2]$ とすると次が成り立つ：$\langle e, m_1\rangle \in LZ[T_1]$, $\langle e, m_2\rangle \in LZ[T_2]$ かつ $m_1 > s \times |L[T_1]|$, $m_2 > s \times |L[T_2]|$．ここで $\langle e, m\rangle \in LZ[T_1 * T_2]$ とすれば，$m = m_1 + m_2 > s \times (|L[T_1]| + |L[T_2]|) = s \times |L[T_1 * T_2]|$ である．これより e は L$[T_1 * T_2]$ で頻出であるから $e \in C[T_1 * T_2]$.
(b) (a) と同様に，$e \in C[T_i]$, $e \in C[T_j]$ が同時に成立しないとき $e \in C[T_i * T_j]$ でないことを示すことができる． □

例 3 定理 2 で等号が成り立たない例を示す．T=$T_1 * T_2 * T_3$ とし，$s = 0.33$ のログ L が次で与えられているとする：

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

T 上の時制頻出クラス C は $\{a, b, c\}$ を含む．$C[T_1] = \{a, c\}$, $C[T_2] = \{b, c\}$, $C[T_3] = \{a, b\}$, $C[T_1 * T_2] = \{c\}$, $C[T_{23}] = \{b\}$ である．したがって $C[T_1 * T_2] = C[T_1] \cup C[T_2]$ ではない．また $C[T_1] \cap C[S]$= $C[T_1 * S]$ ではない． □

時区間を拡張して，隣接したクラスを結合できる場合がある．

定理 3 時区間 T 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．$1 \leq i < n$ とするとき $C[T_i]$ =$C[T_{i+1}]$ ならば $C[T_i] = C[T_i * T_{i+1}]$ である．いい換えると，連続する時区間の間で変化が無ければ隣接時区間を合併しても本質的にかわらない（証明）定理 2 より $C[T_i] \cup C[T_{i+1}] \supseteq C[T_i * T_{i+1}] \supseteq C[T_i] \cap C[T_{i+1}]$ である．$C[T_i]$=$C[T_{i+1}]$ であるから $C[T_i]$=$C[T_i] \cup C[T_{i+1}]$ =$C[T_i] \cap C[T_{i+1}]$ となる． □

後述するように，上記定理の逆は成り立たない．

5. 時制クラスの分割

前章では，どのような時制頻出クラスに対しても，これを記述する部分時制頻出クラス列が存在することを示した．この章では ‘記述する’ ことの意味とクラス列が満たすべき特徴を論じる．

T の分割を $T_1 * .. * T_n$ $(n \geq 1)$，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 $(i = 1, .., n)$．クラス列 $C_1, .., C_n$ が C を ‘記述する’ とは次の 3 つを満たすときと定義する：

(1) C_i は T_i 上の部分時制頻出クラス $(i = 1, 2, .., n)$
(2) $C = C_1 \cup ... \cup C_n$
(3) $C_i \neq C_{i+1}$ ($n > 1$ かつ $i = 1, .., n-1$ のとき)

ここで，C_i は空集合でもよい．

定理 4 T 上の時制頻出クラスを C(T), T = $T_1 * .. * T_n$ とするとき，T = $T'_1 * .. * T'_m$ および T'_i 上のクラス C_i が存在し $(i = 1, .., m)$, $C_1, .., C_m$ は C を記述する．ただし $T_1, .., T_n$ は $T'_1, .., T'_m$ より詳細である．
（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 i があれば定理 3 より T_i, T_{i+1} を合併する．また $C[T_{i+1}]$ は $C[T_i]$ と同じであり，これを取り去る．番号を付け替えて $T'_1, .., T'_m$ を得ることにより，(3) が成り立つ． □

例 4 時制クラス C を記述するクラス列が複数存在する例を示す．次のログを考える．

	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)

(1) 最小サポートを 0.50，ログが (a) で与えられるとする．このとき $C = \{a, b\}$, $C[T_1] = \{a\}$, $C[T_2] = \{b\}$, $C[T_3] = \{a\}$, $C[T_1 * T_2] = \{b\}$, $C[T_2 * T_3] =$

$\{b\}$ である．また $\{C[T_{12}], C[T_3]\}$，$\{C[T_1], C[T_{23}]\}$，$\{C[T_1], C[T_2], C[T_3]\}$ は C を記述する．

(2) 最小サポートを 0.50，ログが (b) とする．このとき $C = \{a, b\}$ である．したがって $C[T_1] = \{a\}$，$C[T_2] = C[T_3] = \{b\}$，$C[T_4] = \{a\}$ である．また $C[T_1 * T_2] = \{a, b\}$，$C[T_2 * T_3] = b$，$C[T_3 * T_4] = \{a, b\}$，$C[T_1 * T_2 * T_3] = \{b\}$，$C[T_2 * T_3 * T_4] = \{b\}$ である．この結果，次のクラス列

$\{C[T_1], C[T_{234}]\}$, $\{C[T_{123}], C[T_4]\}$,
$\{C[T_{12}], C[T_3], C[T_4]\}$,
$\{C[T_1], C[T_{23}], C[T_4]\}$,
$\{C[T_1], C[T_2], C[T_{34}]\}$

は C を記述する．しかし $\{C[T_{12}], C[T_{34}]\}$，$\{C[T_1], C[T_2], C[T_3], C[T_4]\}$ は記述しない．

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

(c)

(3) 最小サポートを 0.50, ログが (c) とする．このとき T 上の時制頻出クラスとして $C = \{a, b\}$ となる．$C[T_1] = C[T_2] = \{a\}$，$C[T_3] = C[T_4] = \{b\}$ であることは容易に分かる．また $C[T_1 * T_2] = \{a\}$，$C[T_2 * T_3] = \{a, b\}$，$C[T_3 * T_4] = \{b\}$，$C[T_1 * T_2 * T_3] = \{a\}$，$C[T_2 * T_3 * T_4] = \{b\}$ である．これよりクラス列

$\{C[T_1], C[T_{234}]\}$，$\{C[T_{123}], C[T_4]\}$，
$\{C[T_{12}], C[T_{34}]\}$，$\{C[T_1], C[T_{23}], C[T_4]\}$，
$\{C[T_1], C[T_2], C[T_{34}]\}$

は C を記述するが，次はいずれもそうではない．

$\{C[T_{12}], C[T_3], C[T_4]\}$
$\{C[T_1], C[T_2], C[T_{34}]\}$
$\{C[T_1], C[T_2], C[T_3], C[T_4]\}$

□

頻出クラス C(T) を記述する時制クラス列は，一般に複数存在する．本研究では，測定可能な時区間で最小のものが存在すると想定し，これに沿って最も精密な変化を反映するものを考える．‘測定可能な時区間で最小のもの’ を最小時区間単位，この大きさを最小ウインドウサイズと呼ぶ☆．このとき次のように ‘適正な列’ を定義する．時区間 T 上のログ L，T の分割を $T_1 * .. * T_n$ とする．また C を L での T 上の時制頻出クラス，C を記述する時制クラス列を $C_1, .., C_n$ とする．このとき $T_i = S_1 * ... * S_k$ のような T_i の任意の分割 $S_1, .., S_k$ に対して，どの j に対しても $C_i[S_j] = C_i[S_{j+1}]$ ならば T_i 上の C_i は C に関して ‘適正である’ という．どの i についても C_i が C に関して適正ならば $C_1, .., C_n$ は適正であるという．

☆ 本稿ではどの最小時区間単位も同一最小ウインドウサイズであるとする．

例 5 適正および不適正な例を示す．

(1) 例 4 (1) において $\{C[T_1], C[T_2], C[T_3]\}$ は C を記述するクラス列であり適正である．同様に例 4 (2) の $\{C[T_1], C[T_{23}], C[T_4]\}$，例 4 (3) の $\{C[T_{12}], C[T_{34}]\}$ も適正である．これ以外は適正ではない．実際，例 4 (3) の $\{C[T_1], C[T_{23}], C[T_4]\}$ は C を記述するが $C[T_{23}]$ のため適正ではない．

	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) 上記 (a) の T=$T_1 * T_2 * T_3$ 上のログを考え $s = 0.40$ とする．このとき $C[T_1] = \{a\}$，$C[T_2] = \{\}$，$C[T_3] = \{b\}$，$C[T_{12}] = \{a\}$，$C[T_{23}] = \{b\}$ である．$C[T_{12}]$ は $C[T_1] \cup C[T_2]$ に，$C[T_{23}]$ は $C[T_2] \cup C[T_3]$ に分割されるので $\{C\}$，$\{C[T_{12}], C[T_3]\}$，$\{C[T_1], C[T_{23}]\}$ のどれも適正でない．

(3) 上記 (b) を T=$T_1 * T_2 * T_3$ 上のログとする．$s = 0.51$ としたとき $C = C[T_{123}] = \{a\}$，$C[T_1] = C[T_2] = \{\}$，$C[T_3] = \{a\}$，$C[T_{12}] = \{\}$，$C[T_{23}] = \{a\}$ である．これより $C[T_{12}] = C[T_1] \cup C[T_2]$，$C[T_{23}] = C[T_2] \cup C[T_3]$ となり $\{C\}$，$\{C[T_1], C[T_{23}]\}$ のどちらも適正ではないが $\{C[T_{12}], C[T_3]\}$ は適正である．

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

(c)

(4) 上記 (c) を T=$T_1 * T_2 * T_3$ 上のログとする．$s = 0.51$ としたとき $C = C[T_{123}] = \{a\}$，$C[T_1] = \{\}, C[T_2] = \{a\}$，$C[T_3] = \{\}$，$C[T_{12}] = \{a\}$，$C[T_{23}] = \{a\}$ である．これより $C[T_{12}]$ は $C[T_1]$，$C[T_2]$ に分解でき，$C[T_{23}]$ は $C[T_2]$，$C[T_3]$ に分解できる．したがって $\{C\}$，$\{C[T_1], C[T_{23}]\}$，$\{C[T_{12}], C[T_3]\}$ のいずれも適正ではない． □

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

（証明）T を最小時区間単位に分割し $T_1, .., T_N$ とする．$C[T_i] = C[T_{i+1}]$ である限り $C[T_i]$ と $C[T_{i+1}]$，および T_i と T_{i+1} を結合する．n は有限であるから，どの隣接時区間もこれ以上結合できなくなるまで繰り返す．必要なら添え字を付け替えて，T の分割 $T_1, .., T_n$ およびその上のクラス列 $C_1, .., C_n$ を得る（ただし $n \leq N$）．これが適正であることをいう．

各 S_i が最小時区間単位なら，定義より分割できないのでクラス列は適正である．$S_i = T_a * .. * T_b$ とする．U_1，U_2 を S_i を得るために，最後に合併した 2 つの時区間とすれば $S_i = U_1 * U_2$ であり，定義より $C[S_i] = C[U_1] = C[U_2]$ である．帰納法より $C[T_a] = ... = C[T_b] = C[S_i]$ であるから，クラス列は適正条件を満たす． □

定理 6 定理 5 において T の分割 $T_1 * .. * T_n$，および T 上で C を記述する適正なクラス列 $C_1, .., C_n$ は一意である．

（証明）$T_1, .., T_n$，$S_1, .., S_m$ をともに T の適正なクラス列を与える時区間列とする．$T_1 = S_1$ で，しかも $n = m$ を証明できれば，2 つの列は帰納法により同一である．

$T_1 = [t_1..t_2)$，$S_1 = [s_1..s_2)$ のとき $T_1 = S_1$ を示す．ともに T の分割における最初の時区間であるから $t_1 = s_1$ であり，したがって $T_1 \subseteq S_1$ または $S_1 \subseteq T_1$ が成り立つ．$T_1 \subseteq S_1$ のとき $t_2 \leq s_2$ である．$t_2 < s_2$ であれば $S_1 = T_1 * U_1$ と分割できる．ただし $U_1 = [t_2..s_2)$．適正な列であることから $C[T_1] = C[U_1] = C[S_1]$ である．ここで，$U_1 \cap T_2 \neq \phi$ なので $C[U_1] = C[T_2]$ $C[T_1] \neq C[T_2]$ とは矛盾する．したがって $t_2 = s_2$ である．

$n = m$ をいう．$n \neq m$ として矛盾を示す．$T_i \neq S_i$ となる最小の i が存在する．$T_i = [t_i..a)$, $S_i = [s_i..b)$ とすれば仮定より i が最小なので，$t_i = s_i$ である．$a \neq b$ であるから $T_i \subseteq S_i$，$a < b$ としてよい．このとき上と同様にして矛盾を得る． □

この結果，時区間の一意分割と適正な時制頻出クラス列を得るためには，次のアルゴリズムを考えることができる．

(1) T 上のログ L を時区間 T とし，w を最小ウインドウサイズとする．
(2) T を w を用いて $T_1, .., T_n$ へ分割する．
(3) $C(T)$ を計算する．また $i = 1, .., n$ に関して $C[T_i]$ を計算する．
(4) $C[T_i] = C[T_{i+1}]$ である限り T_i と T_{i+1} を合併し，添え字を付け直す．
(5) $C[T_1], .., C[T_m]$ $(m \leq n)$ は適正なクラス列となる．

ステップ (2) では 1 項目集合を生成するためにログを 1 度だけ走査する．ステップ (3) は集合一致を行う手順を n 回繰り返すだけの手間が必要である．

6. 実　　験

この章では，本アプローチの有効性を検証するため，気象データを用いた実験を行う．実験に用いたデータは，東京気象台の作成した 1 年分の気象データ 1.03 MB である[17]．この実験では 3 つの属性を対象とする：`date`(00:00-12:31)，`time` (00-24)，`temperature`．ただし実験を容易にするため，空値を取り除き，各 `temperature` 値を `A` から `G` の記号に変換し次のような範囲値で置き換える：

A(-5°)，B(5°-10°)，C(10°-15°)，D(15°-20°)，
E(20°-25°)，F(25°-30°)，G(30°-)

この変換によって，最終的に 8,760 件 94.1 KB のサイズになった．すべてのログは時間順に並べられ最小時区間単位に分割し，提案するアルゴリズムで解析する．本実験では最小ウインドウサイズを 3 日，1 週間，1 カ月とし，最小サポート 5%，15%で評価し，合計 6 種類の分析を行う．

初めに時制頻出クラス `C[00:00:00-12:31:24]` を最小サポート値 0.05，0.15 で抽出する．この結果，それぞれ 6 項目（A，B，C，D，E，F）および 4 項目（B，C，D，E）のデータを得た．

最小サポート 5%で最小ウインドウサイズ 3 日の場合（計 122 時区間）では，最終的に 78 時区間を得た．特に，夏と冬に比較的大きい時区間を得ていることが特徴的である．また，6 月（雨期），9 月から 10 月（台風到来）では観測値の変動が多い．同様に 1 週間ウインドウ（計 52 時区間）の場合は 39 時区間，1 カ月ウインドウ（計 12 時区間）では 9 時区間を得た．1 週間ウインドウの場合についても同様の傾向がより顕著に表れる．しかし，1 カ月ウインドウの場合，あまりに大きな単位であることから，多くの時区間が合併され局所的な特性を失っている．これらの状況を**図 1** に示す．図の各項目に示された値は，併合された最小時区間単位の数である．

これに対して最小サポート 15%の場合，異なる傾向の結果を得た．実際 57 時区間（最小ウインドウサイ

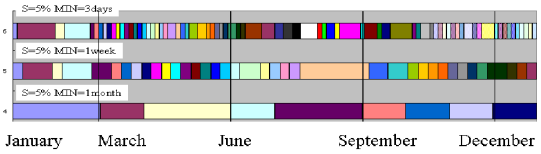


図 1 気象情報の分析（1）：最小サポート 5%
Fig. 1 Analysis of weather data (1): 5% support.

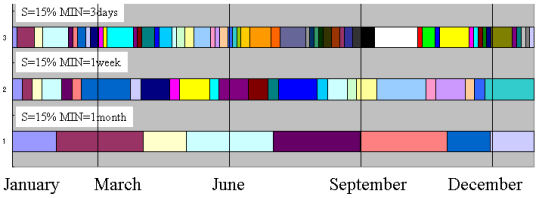


図 2 気象情報の分析（2）：最小サポート 15%
Fig. 2 Analysis of weather data (2): 15% support.

表 1 時区間数
Table 1 Number of intervals.

最小サポート	3 日 (122)	1 週間 (52)	1 カ月 (12)
5%	78(64%)	39(75%)	9(75%)
15%	57(47%)	26(50%)	8(67%)

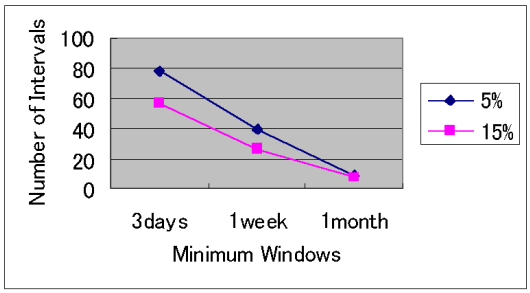


図 3 気象情報の分析（3）：時区間数
Fig. 3 Analysis of weather data (3): intervals.

ズ 3 日），26 時区間（1 週間），8 時区間（1 カ月）となった．これらを**図 2** に示す．最小サポート 5%の場合と比較すると，3 日ウインドウでは冬季や雨期には同様の傾向を読み取ることができる．しかし，これ以外は異なる傾向を示し多くの時区間が合併され局所的な特性を失っている．

時区間数の減少を**表 1** と**図 3** に示す．X 軸は最小ウインドウサイズを，Y 軸は時区間数を示す．最小サポート 5%の場合，時区間数はウインドウサイズの増加につれて著しく減少する．15%の場合も同様の傾向を示すがさほど顕著ではない．

わが国では四季の移り変わりは比較的明確に区別でき，地域によっては完全に観測値の動向が異なる．東京地域の場合，夏季の気温は通常 30 度 C を超えるが，冬季には零度以下となるのも珍しくない．また，雨期は通常 6 月に，台風の到来は 9 月から 10 月の間に生じ観測値が変動することが多い．実験の結果から，時制クラスによってこのような特長的な時区間を抽出できることを示しており，最小サポートおよび最小ウインドウサイズの選択が時区間の合併に大きく影響を与えることが分かる．またおそらくウインドウサイズの設定方法にも依存するであろう．むろん，両パラメータの選定は問題領域に依存するものと考えられるが，雨期や台風到来期では気象に変化を見出す必要があることから，これらのパラメータ値は，サンプリング等の経験から決定する必要があることを示している．

7. 結　　論

本研究では時系列データから時制クラススキーマを発見する手法を提案した．

時間軸（あるいはログ）に沿った時制オブジェクト集合から時間とともに生じる変化を分析し，ある特性を持った時制頻出クラスを得ることができることを示した．頻出時区間を抽出し，隣接する部分クラスが同一オブジェクト集合を有する限り合併する．この結果，一意の時区間分割とそれらに関連した時制クラス列に達することを示した．また実験結果を通じて，このアプローチの有用性を示した．

本研究で示したアプローチは，データマイニングの観点からとらえた（数値分割に関する）一次元データのクラスタリング問題と考えることもできる[11]．また，筆者らはすでに確率的な仮定の下でエルゴード性に基づく技術を論じている[12]．今後はこれらの観点から，本稿で示した手法との融合を図り，時系列データに統合化する方法を展開する予定である．

謝辞 実験等にご助力いただいた渡辺浩平氏（日立製作所）に感謝します．本稿に対し重要な改善点を指摘された査読者の方々にも感謝します．なお，本研究は文部省科学研究費補助金（課題番号 14580392）より一部援助をいただいた．

参 考 文 献

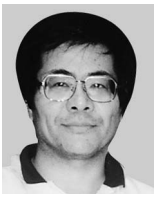
1) Agrawal, R. and Srikant, R.: Fast Algorithms for Mining Association Rules, Proc. *VLDB*, pp.487–499 (1994).
2) Allen, J.: Maintaining Knowledge about Temporal Interval, *CACM* 26-1, pp.832–843 (1983).
3) Antunes, C.M. and Oliveira, A.L.: Temporal Data Mining — An Overview, *KDD Workshop on Temporal Data Mining* (2001).
4) Domingos, P. and Hulten, G.: High Speed Data Streams, *ACM Proc. em KDD* (2000).

5) ElMasri, R., Kouramajian, V., et al.: Temporal Database Modeling, *ACM CIKM* (1993).
6) ElMasri, R. and Navathe, S.: *Fundamentals of Database Systems*, Benjamin/Cummings Pub. (1994).
7) Fayyad, U.M., Piatetsky-Shapiro, G., et al.: *Advances in Knowledge Discovery and Data Mining*, MIT Press (1996).
8) Silva, F.S., Shiel, U. and Catarci, T.: Visual Query Operators for Temporal Databases, *Proc. TIME*, pp.46–54 (1997).
9) Gadia, S.: A Homogeneous Relational and Query Language for Temporal Databases, *ACM TODS* 13-4 (1988).
10) Han, J. and Kamber, M.: *Data Mining*, Morgan Kauffman Pub. (2001).
11) Jain, A.K., Murty, M.N. and Flynn, P.J.: Data Clutering — A Review, *ACM Computing Surveys*, Vol.31, No.3, pp.264–323 (1999).
12) Miura, T. and Shioya, I., et al.: Behavior Discovery as Database Scheme Design, *Proc. TIME*, pp.115–122 (2000).
13) Navathe, S. and Ahmed, R.: A Temporal Relational Model and A Query Language, *Information Sciences*, Vol.49, No.2, pp.147–175 (1989).
14) Savasere, A., Omiecinski, E. and Navathe, S.: An Efficient Algorithm for Mining Accosiation Rules in Large Databases, *Proc. VLDB*, pp.432–444 (1995).
15) Snodgras, R.: The Temporal Query Language TQuel, *ACM TODS*, Vol.12, No.2, pp.247–298 (1987).
16) Tansel, A.U., et al.: *Temporal Databases*, Benjamin/Cummings Pub. (1993).
17) 日本気象協会（編）：気象データひまわり，丸善 (1998).

(平成 14 年 12 月 27 日受付)
(平成 15 年 4 月 10 日採録)

（担当編集委員　定兼 邦彦）

本吉　正博（正会員）
法政大学大学院工学研究科電気工学専攻，現在修士課程在学中．データマイニング，データベース，多変量解析等の分野に興味を持つ．電子情報通信学会会員．



三浦　孝夫（正会員）
京都大学理学部．工学博士（東京大学）．現在，法政大学工学部情報電気電子工学科教授．データモデル，知識表現，演繹データベース，複合オブジェクト等の分野の研究に従事．電子情報通信学会，ACM 各会員．著書に「データモデルとデータベース」（全 2 巻，サイエンス社）．

塩谷　　勇（正会員）
東京電機大学大学院修士課程修了．現在，産能大学経営情報学部教授．時系列モデルの同定，論理プログラミング，グラフ文法，論理データベースの研究に従事．電子情報通信学会，ACM 各会員．