確率的データストリームにおける パターン照合結果の時間的重複に基づくグループ化

杉浦 健人[1,a)]　石川 佳治[1,b)]　佐々木 勇和[2,c)]

概要：データをリアルタイムに解析する複合イベント処理は，日々大量に生成されるデータを素早く活用できるため，学術および商用の両面で広く研究されている．特にパターン照合による複合イベントの検出は，ユーザが求める情報を柔軟に記述・検出できるため，さまざまな応用が提案されている．パターン照合を各イベントに生起確率が付与された確率的データストリームに対して行う場合，データの曖昧性により，同じ時間帯で互いに重複するマッチが検出される．重複して存在するマッチは，いずれもその時間帯にパターンに対応するイベントが生起した可能性を示しており，個々を区別することは必ずしも適当ではない．そこで本稿では，重複したマッチをグループとしてひとまとめにする手法を提案する．まず，本稿で想定する確率的データストリームとパターンの記述方法について述べる．次に，グループの確率的な意味について考えるために，確率的データストリームに対してパターン照合を行った際の確率空間を定義する．その後，階層的クラスタリングにおける単連結法および完全連結法を参考に，グループを生成するためのセマンティクスとして単オーバラップと完全オーバラップを定義する．また，マッチをグループにまとめるアルゴリズムと，トランスデューサを用いてグループの確率を効率的に計算する手法を提案する．最後に，評価実験により本手法の有効性を示す．

1. はじめに

データをリアルタイムに解析する複合イベント処理は，日々大量に生成されるデータを素早く活用できるため，学術および商用の両面で広く研究されている．特に**パターン照合**による複合イベントの検出は，ユーザが求める情報を柔軟に記述・検出できるため，株価の急激な変動の検出などさまざまな応用が提案されている [4]．しかし，パターン照合に関する既存研究の多くは入力となる情報の曖昧性を考慮していない．情報源がセンサ機器などである場合，センサ自体の計測誤差や情報伝達時の欠落によって，得られる情報は曖昧になる．このような曖昧さは確率によって表せるため，入力される情報は図 1 のような**確率的データストリーム**となる．

確率的データストリームに対するパターン照合には，ユーザに冗長な情報を出力してしまうという問題がある．例えば，図 2 はパターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+\rangle$ に対応するマッチを図 1 から検出したものであるが，いずれのマッチもこの時間帯でパターンに対応するイベントが生起したことを示している．しかし図 1 を見る限り，パターンに対応するイベントは現実には一度しか生起していないと考える方が自然である．つまり図 2 は，現実には一つしか存在しないマッチが情報の曖昧さにより複数のマッチとなって出力されていることを示している．こうしたマッチをすべて確認するのはユーザにとって煩雑であるため，既存手法では生起確率の小さいマッチを重要度の小さいマッチとし，そのようなものを除くことでユーザへ返すマッチを制限している [3]．

しかし，いずれのマッチもその時間帯でパターンに該当するイベントが生起したという情報を持つため，単純に冗長なマッチを削除するだけではマッチが持つ情報，つまりパ

時刻		1	2	3	4	5	6	7
イベント	a	1.0	0.3	0.1	0.2	0	0	0.3
	b	0	0.7	0.7	0.7	0.9	0	0.7
	c	0	0	0.2	0.1	0.1	1.0	0

図 1　確率的データストリーム

マッチ	時刻						確率
	1	2	3	4	5	6	
m_1	a	b	c				0.14
m_2	a	b	b	b	b	c	0.3087
m_3	a	a	b	b	b	c	0.1323
m_4		a	b	b	b	c	0.1323
m_5				a	b	c	0.18

図 2　パターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+\rangle$ に対するパターン照合結果の一部

[1] 名古屋大学大学院情報科学研究科
[2] 名古屋大学未来社会創造機構
[a)] sugiura@db.ss.is.nagoya-u.ac.jp
[b)] ishikawa@is.nagoya-u.ac.jp
[c)] yuya@db.ss.is.nagoya-u.ac.jp

ターンに対応するイベントが生起した確率まで削除してしまう．そこで，本稿では同じイベントの生起を示すマッチを一つの**グループ**にまとめ，出力の冗長さを軽減するとともにマッチの生起確率を集約する手法を提案する．例えば図 2 のマッチの場合，グループ $g = \{m_1, m_2, m_3, m_4, m_5\}$ にまとめ，生起確率を $P(g) = 0.7358$ に集約する．本手法により，マッチの持つ情報を損なうことなく，出力の冗長さの軽減が可能となる．

2. 準備

確率空間の定義．確率的データストリーム上でのパターン照合における確率空間を定義する．まず，確率的データストリームの要素となる**確率的イベント**を以下に示す．

定義 1　**確率的イベント** e_t は，イベントの各属性値 $\alpha \in D$ に対して生起確率 $P(e_t = \alpha)$ を持つ時刻 t のイベントである．ただし，D はイベントの離散的なドメインである．また，イベントの生起確率 $P(e_t = \alpha)$ は以下の式を満たす．

$$\forall \alpha \in D,\ 0 \leq P(e_t = \alpha) \leq 1$$

$$\sum_{\alpha \in D} P(e_t = \alpha) = 1 \qquad \square$$

確率的イベントを用いて**確率的データストリーム**を以下のように定義する．

定義 2　**確率的データストリーム** $PDS = \langle e_i, e_{i+1}, ..., e_j, ... \rangle$ は，確率的イベントの系列である．$\square$

例えば，図 1 の確率的データストリームは，ドメインが $D = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ である確率的イベント e_t の系列 $PDS = \langle e_1, e_2, e_3, e_4, e_5, e_6, e_7 \rangle$ として表せる．なおこれ以降，時刻 t にイベントが $\alpha \in D$ であることを α_t で表す．

次に，時刻 t において生起確率が 0 より大きい属性の集合を D_t で表し，$D_t \times D_{t+1}$ で時刻 t と $t+1$ におけるイベントの系列の全集合を表すとして，確率的データストリームにおける可能世界を以下のように定義する．なお，$[x : y]$ は開区間を，$(x : y)$ は閉区間を表す．

定義 3　有限長の確率的データストリーム $PDS = \langle e_i, e_{i+1}, ..., e_j \rangle$ が与えられたとき，PDS の可能世界 w の全集合は $W_{[i:j]} = D_i \times D_{i+1} \times ... \times D_j$ である．また，可能世界 $w = \langle \alpha_i, \alpha_{i+1}, ..., \alpha_j \rangle$ に対する確率は，$P(w) = \prod_{\alpha_t \in w} P(\alpha_t)$ で与える．$\square$

例えば図 1 の確率的データストリームの時区間 $[2 : 3]$ について考えると，$W_{[2:3]} = \{\mathtt{a}_2, \mathtt{b}_2\} \times \{\mathtt{a}_3, \mathtt{b}_3, \mathtt{c}_3\} = \{\langle \mathtt{a}_2, \mathtt{a}_3 \rangle, \langle \mathtt{a}_2, \mathtt{b}_3 \rangle, \langle \mathtt{a}_2, \mathtt{c}_3 \rangle, \langle \mathtt{b}_2, \mathtt{a}_3 \rangle, \langle \mathtt{b}_2, \mathtt{b}_3 \rangle, \langle \mathtt{b}_2, \mathtt{c}_3 \rangle\}$ である．また，可能世界 $\langle \mathtt{a}_2, \mathtt{b}_3 \rangle$ の生起確率は $P(\langle \mathtt{a}_2, \mathtt{b}_3 \rangle) = P(\mathtt{a}_2) \times P(\mathtt{b}_3) = 0.21$ である．

上述した可能世界を用いて，確率的データストリームにおける確率空間を次のように定義する．

定義 4　有限長の確率的データストリーム $PDS = \langle e_i, e_{i+1}, ..., e_j \rangle$ に対する確率空間は $(2^{W_{[i:j]}}, P)$ である．ただし，$2^{W_{[i:j]}}$ は $W_{[i:j]}$ のべき集合を，P は $x \in 2^{W_{[i:j]}}$ に対して $P(x) = \sum_{w \in x} P(w)$ で確率を与える関数である．$\square$

問合せパターンとマッチの定義．問合せパターンの記述方法を定める．本稿では，人の行動・移動などイベントの遷移が連続的なものを入力として想定する．つまり，$\langle \mathtt{a}_1, \mathtt{a}_2, \mathtt{b}_3, \mathtt{b}_4, \mathtt{b}_5 \rangle$ などのように，イベントが常に複数の時刻にまたがって存在する状態を想定する．このような状態を想定すると，$\langle \mathtt{a\ b} \rangle$ のように単体のイベントのみを検出するパターンの有用性は小さいと考えられる．そこで，$\langle \mathtt{a}^+\mathtt{b}^+ \rangle$ のように，問合せパターンではすべてのイベントに対してクリーネ閉包を付与する．また，クリーネ閉包の付与を前提とするため，$\langle \mathtt{a}^+\mathtt{a}^+\mathtt{b}^+ \rangle$ のような同じイベントの連続した記述は許可しない．なお，本稿では選言や否定など，パターンに対するその他のオプションの使用は考慮しない．

次に，マッチとその生起確率の定義を以下に示す．

定義 5　マッチ m はユーザが指定したパターンに対応するイベントの系列である．確率空間 $(2^{W_{[i:j]}}, P)$ に対して，m を部分系列に含む可能世界の集合を $W_m \subseteq W_{[i:j]}$ とするとき，マッチの確率は $P(m) = \sum_{w \in W_m} P(w)$ である．$\square$

例えば，図 1 における $W_{[1:4]}$ の可能世界で考えると，図 2 の m_1 の確率は以下の可能世界の確率の和となる．

$$w_1 = \langle \mathtt{a}_1, \mathtt{b}_2, \mathtt{c}_3, \mathtt{a}_4 \rangle,\ P(w_1) = 0.028$$
$$w_2 = \langle \mathtt{a}_1, \mathtt{b}_2, \mathtt{c}_3, \mathtt{b}_4 \rangle,\ P(w_2) = 0.098$$
$$w_3 = \langle \mathtt{a}_1, \mathtt{b}_2, \mathtt{c}_3, \mathtt{c}_4 \rangle,\ P(w_3) = 0.014$$

つまり，$P(m_1) = P(w_1) + P(w_2) + P(w_3) = 0.14$ である．

3. グループの定義

複数のマッチを一つのグループにまとめるとき，それらが同じイベントの生起に対応していることをどのような基準で判断するかが重要となる．本稿では，生起した時間が近いマッチを同じイベントに対応するマッチと考え，マッチ同士の時間的なオーバラップに注目してグループ化を行う．なお，マッチが時間的にオーバラップするとは，図 2 における m_1, m_2 の時刻 2, 3 のように，二つのマッチがある時刻で互いに何らかのイベントを持つことを示す．これ以降，$ts\_overlap(m_i, m_j)$ を m_i, m_j の時間的なオーバラップを示す述語記号として扱う．

マッチの時間的なオーバラップを用いて，本稿では階層的クラスタリングにおける単連結法 (single-link method) と完全連結法 (complete-link method) の考えを参考にグループとなるマッチの集合を定義する [2]．単連結法は，あるクラスタ内に存在するすべてのドキュメントが同じクラスタ内に少なくとも一つ類似なものを持つようクラスタリングする手法である．一方，完全連結法では，あるクラスタ内に存在するすべてのドキュメントが互いに類似となるようクラスタを生成する．以下では，各手法を参考にした単オーバラップと完全オーバラップについて順に説明し，最後にグループの生起確率を定義する．

単オーバラップに基づく定義．単オーバラップの定義を以下に示す．

定義 6　マッチの集合 M が以下の式を満たすとき，M は**単オーバラップ**の性質を持つ．

$$\forall m_i \in M, \exists m_j \in M, m_i \neq m_j \wedge ts\_overlap(m_i, m_j) \quad \square$$

単オーバラップは，グループ内のマッチが他のいずれかのマッチとオーバラップすることを保証する．例えば図 2 の場合，m_2, m_3, m_4 がすべてのマッチとオーバラップするため，五つのマッチすべてを含んだグループが生成される．

単オーバラップにより生成されるグループは，言い換えれば，ある時区間に存在するマッチをすべてひとまとめにしたものである．先ほどの例の場合，生成されたグループは時区間 $[1:6]$ に存在するマッチをすべてひとまとめにしたものとみなせる．そこで，単オーバラップに基づくグループを以下のように定義する．

定義 7　グループは単オーバラップの性質を持つマッチの集合である．グループは $g = (t_s, t_e)$ のように，グループの開始時刻，終了時刻の組で表す．　$\square$

つまり，単オーバラップによるグループ $g = (t_s, t_e)$ は，時区間 $[t_s, t_e]$ の範囲内でパターンに対応するイベントが $P(g)$ の確率で生起したことを示す．

完全オーバラップに基づく定義．完全オーバラップの定義を以下に示す．

定義 8　マッチの集合 M が以下の式を満たすとき，M は**完全オーバラップ**の性質を持つ．

$$\forall m_i, m_j \in M, ts\_overlap(m_i, m_j) \quad \square$$

完全オーバラップは，グループ内のすべてのマッチが互いにオーバラップすることを保証する．例えば図 2 の場合，m_1, m_5 が互いにオーバラップしないため，$g_1 = \{m_1, m_2, m_3, m_4\}$ と $g_2 = \{m_2, m_3, m_4, m_5\}$ という二つのグループが生成される．

完全オーバラップによるグループは，単オーバラップと違い開始時刻と終了時刻のみではグループを区別できない．しかし，グループ内で初めてマッチを受理した時刻を用いることで区別できる．先ほどの例の場合，各グループの開始時刻と終了時刻は同じであるが，g_1 は時刻 3 に，g_2 は時刻 6 に初めてマッチを受理したという点で異なる．これは，マッチを初めて受理した時刻がそのグループを構成するマッチの開始時刻と終了時刻の切れ目を示しているためである．上述の例であれば，g_1 が時刻 3 に初めてマッチを受理したことが，このグループを構成するマッチが $[1:3]$ の区間で開始し $[3:6]$ の区間で終了することを示している．したがって，完全オーバラップに基づくグループを以下のように定義する．

定義 9　グループは完全オーバラップの性質を持つマッチの集合である．グループは $g = (t_s, t_f, t_e)$ のように，グループの開始時刻，マッチの初受理時刻，グループの終了時刻の組で表す．　$\square$

Input: PDS ▷ 確率的データストリーム
$G \leftarrow \emptyset$ ▷ グループの候補の集合
$R \leftarrow \emptyset$ ▷ マッチの候補の集合
$M_t \leftarrow \emptyset$ ▷ 時刻 t に受理したマッチの集合
for all $e_t \in PDS$ **do**
　イベント e_t を用いて R, M_t を更新
　if $M_t \neq \emptyset$ **then** $updateGroupsForSingle(G, M_t)$
　for all $g \in G$ **do** ▷ グループの出力処理
　　if $\nexists r \in R, \forall m \in g, ts\_overlap(r, m)$ **then**
　　$P(g)$ を計算し g を出力
　　$G \leftarrow G \setminus \{g\}$

procedure $updateGroupsForSingle(G, M_t)$
　$G \leftarrow G \cup \{M_t\}$ ▷ M_t を新しいグループとして追加
　for all $g_i, g_j \in G\ (i \neq j)$ **do**
　　if $\exists m_i \in g_i, \exists m_j \in g_j, ts\_overlap(m_i, m_j)$ **then**
　　$g_i \leftarrow g_i \cup g_j$
　　$G \leftarrow G \setminus \{g_j\}$

図 3　単オーバラップに基づくグループの生成アルゴリズム

すなわち，完全オーバラップによるグループ $g = (t_s, t_f, t_e)$ は，パターンに対応するイベントが時区間 $[t_s : t_f]$ の間で始まり $[t_f : t_e]$ の間で終了したことを示す．単オーバラップにより生成されたグループがある範囲内にパターンが存在するかしないかのみを示していたのに対し，完全オーバラップによるグループではイベントの開始・終了時刻のおおよその範囲を示している．これにより，単オーバラップよりも詳細にイベントが検出できる．

グループの生起確率の定義．定義 5 に示した通り，マッチの生起確率はマッチを部分系列に含む可能世界の確率の和である．したがって，マッチの集合であるグループの生起確率は，各マッチを部分系列に含むすべての可能世界の確率の和として考えるのが自然である．グループの生起確率の定義を以下に示す．

定義 10　グループ g の確率を以下の式で与える．なお，$W_g = \bigcup_{m_i \in g} W_{m_i}$ であり，W_{m_i} は m_i を部分系列に含む可能世界の集合を示す．

$$P(g) = \sum_{w \in W_g} P(w) \quad \square$$

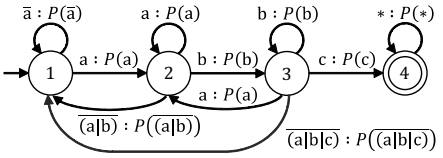
4. グループの生成アルゴリズム

本章ではグループの生成方法について述べる．まず単オーバラップを用いる場合について説明し，次に完全オーバラップの場合を述べる．なお，本稿ではマッチの検出方法については議論せず，各時刻 t でその時刻に受理するマッチの集合 M_t が得られると想定する．ただし，検出されるマッチの数は指数関数的に増加するため，マッチの生起確率にしきい値 θ を与え，θ より大きい生起確率を持つマッチのみ検出する．

単オーバラップの場合．単オーバラップに基づくグループの生成アルゴリズムを図 3 に示す．ある時刻 t で受理されたマッチの集合 M_t は，時刻 t ですべてのマッチが必ずオーバラップするため，それ自体を一つのグループとみな

```
1: procedure updateGroupsForComplete(G, M_t)
2:     for all g ∈ G do
3:         for all m ∈ M_t do
4:             if ∀m' ∈ g, ts_overlap(m, m') then
5:                 g ← g ∪ {m}
6:     G ← G ∪ {M_t}                ▷ M_t を新しいグループとして追加
```

図 4 完全オーバラップに基づくグループの生成アルゴリズム



∗ : 任意のイベント

図 5 パターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+\rangle$ のトランスデューサ

せる．また，単オーバラップの条件により，異なるグループ間でオーバラップするマッチが一つでも存在すれば，それらは一つのグループにまとめられる．そこで，各時刻 t で M_t をグループの候補として生成し，その後オーバラップするマッチを持つグループを結合していくことで単オーバラップによるグループの生成を行う．

例として，図 1 の確率的データストリームにしきい値 0.1，パターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+\rangle$ を与えた際の単オーバラップに基づくグループの生成過程を説明する．なお説明において，m_i は図 2 の各マッチと対応する．時刻 1, 2 では $M_1 = M_2 = \emptyset$ であるためほとんどの処理は行われず，マッチの候補として $R = \{(\langle \mathtt{a}_1, \mathtt{a}_2\rangle, 0.3), (\langle \mathtt{a}_1, \mathtt{b}_2\rangle, 0.7), (\langle \mathtt{a}_2\rangle, 0.3)\}$ が生成される（5 行目）．時刻 3 では，マッチが受理され $M_3 = \{m_1\}$ となるため，M_3 をグループの候補 g_1 として G に加える（13 行目）．この段階で $R = \{(\langle \mathtt{a}_1, \mathtt{a}_2, \mathtt{b}_3\rangle, 0.21), (\langle \mathtt{a}_1, \mathtt{b}_2, \mathtt{b}_3\rangle, 0.49), (\langle \mathtt{a}_2, \mathtt{b}_3\rangle, 0.21)\}$ であり，g_1 中のマッチ m_1 とオーバラップするものが存在する（8 行目）．つまり，まだ g_1 にマッチが追加される可能性があるため，グループの出力は行われない．なお，生起確率がしきい値 $\theta = 0.1$ 以下であるマッチの候補が削除されている点に注意する．その後，時刻 4, 5 では $M_4 = M_5 = \emptyset$ であるためグループ化の処理は行われず，マッチの候補 R のみ更新される（5 行目）．時刻 6 では，$M_6 = \{m_2, m_3, m_4, m_5\}$ が候補 g_2 としてグループ G に一度加えられるが，g_2 は $g_1 = \{m_1\}$ とオーバラップするマッチを持つため，g_1 に結合された後削除される（13〜17 行目）．その後，時刻 7 で $R = \{(\langle \mathtt{a}_7\rangle, 0.3)\}$ となり，R 中から $g_1 = \{m_1, m_2, m_3, m_4, m_5\}$ とオーバラップするものが無くなるため，生起確率 $P(g_1)$ を計算した後 $g_1 = (1, 6)$ として出力する（ 8 〜11 行目）．

完全オーバラップの場合． 完全オーバラップに基づくグループ化のための手続きを図 4 に示す．図 3 の 6 行目でこの手続きを呼び出すことで，完全オーバラップに基づくグループが生成できる．完全オーバラップに基づくグループは，3 で述べたように，マッチを初めて受理した時刻によって区別できる．そこで，時刻 t に受理したマッチの集合 M_t を初受理時刻 t のグループ g とし，g に完全オーバラップの性質を満たすマッチを追加していくことでグループの生成を行う．

例として，図 1 の確率的データストリームにしきい値 0.1，パターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+\rangle$ を与えた際の完全オーバラップに基づくグループの生成過程を説明する．ただし，時刻 6 までの処理は単オーバラップの場合と同様であるため省略する．時刻 6 では，g_1 中のマッチ m_1 とオーバラップするマッチが $M_6 = \{m_2, m_3, m_4, m_5\}$ から g_1 に追加される（図 4 の 2〜5 行目）．この場合，m_5 以外のマッチが m_1 とオーバラップするため，$g_1 = \{m_1, m_2, m_3, m_4\}$ となる．また，M_6 を初受理時刻 6 のグループ g_2 として G に追加する（図 4 の 6 行目）．その後，時刻 7 で $R = \{(\langle \mathtt{a}_7\rangle, 0.3)\}$ となり，R 中から g_1, g_2 とオーバラップするものが無くなるため，それぞれの生起確率を計算し $g_1 = (1, 3, 6), g_2 = (1, 6, 6)$ として出力する（図 3 の 8〜11 行目）．

5. グループの生起確率の効率的な計算

3 で述べたように，グループの生起確率はマッチを部分系列として含む可能世界をすべて列挙することで計算できる．しかし，可能世界の数はグループの時間幅（$|g| = t_e - t_s + 1$）に対して指数関数的に増加するため，可能世界を列挙する単純な手法では効率が悪い．そこで，**トランスデューサ**を用いた効率的な計算手法を提案する．

トランスデューサは入力の受理の確認と同時に出力を求めるオートマトンである．例えば，図 5 はパターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+\rangle$ で生成したグループの確率を計算するためのトランスデューサである．このトランスデューサはマッチを部分系列に持つ可能世界をすべて受理するよう作成されており，同時に各可能世界の確率を計算する．つまりこのトランスデューサを用いることで，マッチを部分系列に含む可能世界の確率の総和を計算するという問題を，トランスデューサの受理状態に到達する確率を求めるという問題に変換できる．

単オーバラップの場合． 単オーバラップによって生成されたグループ $g = (t_s, t_e)$ は，時区間 $[t_s : t_e]$ の間に存在するマッチをまとめたものである．したがって，グループの生起確率 $P(g)$ は，時区間 $[t_s : t_e]$ でトランスデューサの受理状態に到達する確率となる．

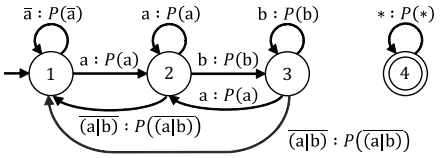
トランスデューサの受理状態に到達する確率は，遷移行列を用いて計算する．例えば，次の式は図 5 のトランスデューサに対応する遷移行列である．

$$T_{t-1,t} = \begin{bmatrix} P(\overline{\mathtt{a}_t}) & P(\mathtt{a}_t) & 0 & 0 \\ P\left(\overline{\mathtt{a}|\mathtt{b}_t}\right) & P(\mathtt{a}_t) & P(\mathtt{b}_t) & 0 \\ P\left(\overline{\mathtt{a}|\mathtt{b}|\mathtt{c}_t}\right) & P(\mathtt{a}_t) & P(\mathtt{b}_t) & P(\mathtt{c}_t) \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

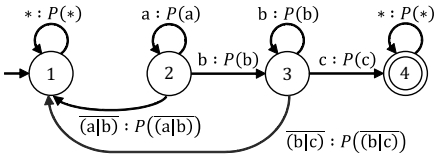
各行が時刻 $t-1$ におけるトランスデューサの状態を，各

時刻	0	1	2	3	4	5	6
$V_t[1]$	1.0	0	0	0.06	0.06	0.08	0.08
$V_t[2]$	0	1.0	0.3	0.10	0.17	0	0
$V_t[3]$	0	0	0.7	0.70	0.56	0.65	0
$V_t[4]$	0	0	0	0.14	0.21	0.27	0.92

図 6 ベクトル V_t の更新（小数点第二位まで表示）



(a) マッチ受理前に使用するトランスデューサ



(b) マッチ受理後に使用するトランスデューサ

図 7 追加で使用するトランスデューサ

列が時刻 t で遷移した先の状態を表す．時刻 t にトランスデューサの各状態に到達する確率をベクトル V_t で表すとき，V_t の確率は 1 単位時刻前の確率 V_{t-1} と遷移行列 $T_{t-1,t}$ を用いて次の式で計算できる．

$$V_t = V_{t-1} \times T_{t-1,t}$$

したがって，時区間 $[t_s : t_e]$ において最終的にトランスデューサの各状態に到達する確率 V_{t_e} は次の式で求められる．ただし，V_{t_s-1} は $V_{t_s-1}[0] = 1$，$V_{t_s-1}[i] = 0\ (i > 0)$ で初期化する．

$$V_{t_e} = V_{t_s-1} \times \prod_{t=t_s}^{t_e} T_{t-1,t}$$

グループの生起確率はこのベクトル V_{t_e} のうち，受理状態にあたる要素の値となる．

実際に，図 1 の確率的データストリームに対してパターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+ \rangle$ で生成したグループ $g = (1, 6)$ の確率を計算する様子を図 6 に示す．このグループの終了時刻は 6 であるため，グループの生起確率 $P(g)$ は約 0.92 である．

完全オーバラップの場合． 完全オーバラップの場合，グループ $g = (t_s, t_f, t_e)$ を構成するマッチは開始時刻が $[t_s : t_f]$，終了時刻が $[t_f : t_e]$ の間にある．しかし，図 5 のトランスデューサは $[t_s : t_f)$ 及び $(t_f : t_e]$ の区間に存在するマッチも受理してしまうため，このままではグループの生起確率が計算出来ない．そこで，図 7 のように各区間のマッチを受理しないトランスデューサを追加で使用することで，完全オーバラップによるグループの生起確率を計算する．(a) はグループの開始時刻 t_s から初受理時刻 t_f の直前まで使用するトランスデューサで，$[t_s : t_f)$ に存在するマッチを受理しないよう図 5 のトランスデューサから受理状態に移る遷移を除くことで作成できる．一方，(b) は初受理時刻 t_f の直後から終了時刻 t_e まで使用するトランスデューサで，図 5 のトランスデューサから新しいマッチの候補を生成する遷移を除くことで，$(t_f : t_e]$ に存在するマッチを受理しないようにしている．

基本的な確率の計算方法は単オーバラップの場合と同じである．ただし，上述の通りグループの時刻に応じて使用するトランスデューサを変更する．図 5 のトランスデューサの遷移行列を T，図 7 の (a) と (b) の遷移行列をそれぞれ T^a, T^b とする．このとき，グループ $g = (t_s, t_f, t_e)$ の確率ベクトル V_{t_e} は以下の式で計算できる．なお，V_{t_s-1} は単オーバラップの場合と同様に初期化する．

$$V_{t_e} = V_{t_s-1} \times \prod_{t=t_s}^{t_f-1} T^a_{t-1,t} \times T_{t_f-1,t_f} \times \prod_{t=t_f+1}^{t_e} T^b_{t-1,t}$$

6. 評価実験

本章では実験により提案手法の有効性を評価する．実験で使用するシステムは，SASE プロジェクトで作成されたシステム SASE+ [1] を拡張することで実装した．

実験で使用したデータセットについて説明する．実験は実データと人工データの両方を用いて行った．実データには，Lahar プロジェクトで公開されている屋内位置の確率的データストリームを使用した．このデータは屋内構造をグラフで表しており，被験者が各ノードにいる確率と実際にその被験者がいた正解ノードの情報を一秒毎に持つ．被験者はある部屋への入退出と廊下の移動を繰り返し行うよう指定されており，合計 9 回部屋への入退出を行っている．そのため，実験ではこの部屋の入退出を $\langle \mathtt{Door}^+\ \mathtt{Room}^+\ \mathtt{Door}^+ \rangle$ として表し，入力パターンとした．人工データには以下の手順で生成したイベント数 100,000 の確率的データストリームを使用する．

(1) 非確率的データストリーム $\langle \alpha_1, ...\alpha_{100000} \rangle$ を生成 $(\alpha_t \in \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\})$.
(2) 各時刻 t で生起しなかったイベント $(\forall \alpha'_t \in \{\mathtt{a}, \mathtt{b}, \mathtt{c}, \mathtt{d}\} \setminus \{\alpha_t\})$ に対して，$[0, 0.1]$ の範囲でランダムに生起確率を付与.
(3) 残りの生起確率 $\left(1 - \sum_{\alpha'_t \in \{\mathtt{a},\mathtt{b},\mathtt{c},\mathtt{d}\} \setminus \{\alpha_t\}} P(\alpha'_t)\right)$ を α_t の生起確率とする.

なお手順 1 の段階で，入力として与えるパターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+ \rangle$ に対応するマッチが生起するようストリームを生成する．

6.1 グループ化の評価

提案手法によって生成されたグループを評価する．この実験では入力に実データとパターン $\langle \mathtt{Door}^+\ \mathtt{Room}^+\ \mathtt{Door}^+ \rangle$ を用いる．

まず，出力されたマッチとグループの数を表 1 に示す．結果から，マッチのグループ化により出力数が大幅に削減

表 1　マッチもしくはグループの出力数

グループ化の方法	しきい値			
	10^{-2}	10^{-3}	10^{-4}	10^{-5}
グループ化なし	677	4,273	13,190	29,543
単オーバラップ	12	14	10	10
完全オーバラップ	180	401	432	448

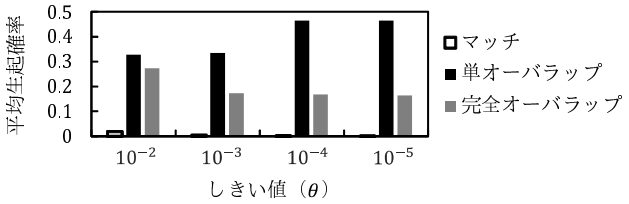


図 8　マッチとグループの生起確率の平均

できていることがわかる．特にしきい値を小さくしたとき，マッチの出力数が大幅に増加するのに対し，グループの出力数がそれほど大きくならない点に注目する．これは検出されるマッチのほとんどがオーバラップすること，つまりマッチによる出力の冗長さが大きいことを示している．対してグループの出力数から，本手法で用いたグループ化が冗長さの削減に有効であることが確認できる．

ただし，しきい値を下げることで単オーバラップに基づくグループの出力数が下がっている点に注意する．これは，しきい値を下げ検出されるマッチを増やしたことで，それまで別々に検出されていたグループが一つにまとめられたことを示す．このようなグループ同士の結合は本来別々に出力すべきグループに対しても行われる可能性があるため，単オーバラップの使用の際には適切なしきい値を選択する必要がある．一方，完全オーバラップに基づくグループにはこのような問題は存在しない．しかし，元々の入退出が 9 回しか行われていない点を考慮すると，完全オーバラップではまだ出力に冗長さが残っていると考えられる．そのため，グループの生成アルゴリズムを改善し，適当なグループのみの出力を行う必要がある．

次に，検出されたマッチとグループの生起確率の平均を図 8 に示す．結果から，マッチをグループにまとめることで生起確率が大幅に上昇していることがわかる．一つ一つのマッチの生起確率は小さいが，表 1 からわかるようにグループは大量のマッチで構成されているため，このような生起確率の向上が可能となっている．なお，しきい値 10^{-3} から 10^{-4} にかけて単オーバラップに基づくグループの生起確率が上昇しているのは，上述したようにグループ同士の結合が起きているためである．

6.2　トランスデューサによる確率計算の評価

本稿で提案したトランスデューサによる生起確率の計算手法を評価する．実験では入力に人工データとパターン $\langle \mathtt{a}^+\mathtt{b}^+\mathtt{c}^+ \rangle$ を使用する．また，比較手法としてマッチを含む可能世界をすべて列挙する単純な手法を使用する．

グループの時間幅に対するグループの生起確率の計算時

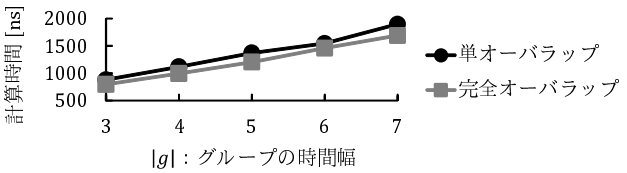


図 9　提案手法によるグループの生起確率の計算時間

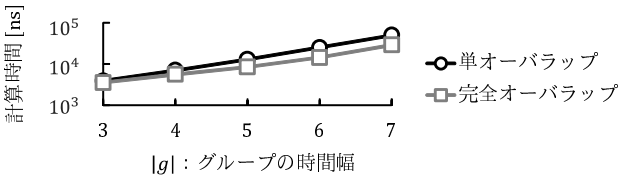


図 10　単純な手法によるグループの生起確率の計算時間

間を図 9，10 に示す．なお，図 10 のみ縦軸が対数軸になっている点に注意する．結果から，提案手法の計算時間が線形に増加しているのに対し，単純な手法は指数関数的に増加していることがわかる．単純な手法では，生起確率の計算に必要な可能世界の個数がグループの時間幅に対して指数関数的に増加するため，計算時間も指数関数的に増える．一方提案手法では，グループの時間幅の大きさは行列の乗算を行う回数となるため，計算時間の増加は線形で済む．

7.　おわりに

本稿では，確率的データストリームでのパターン照合におけるマッチのグループ化手法について提案した．マッチをグループにまとめるための指針として完全オーバラップと単オーバラップを定義し，これらを用いたグループの定義を提案した．また，各オーバラップを使用した場合のグループの生成アルゴリズムと，トランスデューサを用いることで効率良くグループの生起確率を計算する手法を提案した．加えて，人工データと実データによる評価実験を行い，本手法の有効性を確認した．今後の課題としては，グループにまとめるための新しい指針の考案や，異なる時刻のイベントの生起確率が相関を持つ確率的データストリームへの拡張，より詳細な問合せ条件の記述ができる言語の考案などが挙げられる．

謝辞　本研究の一部は科研費（25280039, 2650043）および JST「革新的イノベーションプログラム」による．

参考文献

[1] Agrawal, J., Diao, Y., Gyllstrom, D. and Immerman, N.: Efficient Pattern Matching over Event Streams, *Proc. ACM SIGMOD*, pp. 147–160 (2008).
[2] Jain, A. K., Murty, M. N. and Flynn, P. J.: Data Clustering: A Review, *ACM Comput. Surv.*, Vol. 31, No. 3, pp. 264–323 (1999).
[3] Li, Z., Ge, T. and Chen, C. X.: ε-Matching: Event Processing over Noisy Sequences in Real Time, *Proc. ACM SIGMOD*, pp. 601–612 (2013).
[4] Wu, E., Diao, Y. and Rizvi, S.: High-performance Complex Event Processing over Streams, *Proc. ACM SIGMOD*, pp. 407–418 (2006).