Vol.2017-IFAT-126 No.1
2017/3/30

順序マイニングを用いた系列分類

後藤 仁[1,a)]　白井 匡人[2,b)]　三浦 孝夫[1]

概要：本研究では順序マイニングによって系列中の頻出部分列（エピソード）を獲得し，系列分類に用いる．系列分類は心拍データからの病理診断や文書分類，DNA の遺伝子領域判別など広い応用と需要を持つ．しかしながら判別分析やサポートベクターマシンといったこれまでの手法は系列の各値が状態に対応すると仮定していない．このため系列の持つ状態を手がかりとして分類に用いることができない．本研究では，系列に対応する状態列を扱える条件付き確率場（CRF) による状態遷移モデルをクラス毎に構築し，最尤原理に基づいて分類を行う．本研究では CRF の素性と順序マイニングを対応付けるために，エピソードで系列を書き換える手法を提案する．

キーワード：系列分類, 条件付き確率場, 順序マイニング, エピソード, 素性選択

1. 前書き

近年，ストレージや計測機器の発達によりこれまで捨てられてきた多種多様なデータが記録・収集されるようになっている．そのためこれら膨大なデータを分析し活用することが求められている．中でも系列（値の並び）の分類は遺伝情報の解析やネットワークの侵入検知など幅広い分野で応用を持つ．Xing ら [6] によれば，系列間距離に基づく手法や素性ベクトルに基づく手法，モデルに基づく手法などが系列分類に用いられている．しかし，これらの手法は系列のみからクラスを決定している点で共通している．このため系列各要素に対応する状態を分類に用いることができない．

例えばクラシック楽曲の形式は曲の構成と密接な関連を持つ．ソナタ形式のクラシック楽曲は提示部・展開部・再現部・結尾部という A-B-A'-C の構成をとる．対して三部形式は A-B-A の構成からなる．楽曲の形式判定を行うことを考えると，旋律そのものではなく旋律の各部の構成が大きな手がかりとなるといえる．本研究では曲の構成にあたるものを系列要素に対応する状態の列と捉え，その情報を分類に用いる．分類のために，系列の状態を扱える確率モデルをクラス毎に構築する．そして最尤原理に基づき最も高い尤度で正解状態列を与えるモデルを選択して分類を行う．

モデルが分類の手がかりとする素性には系列の頻出部分列（エピソード）を用いる．これは，エピソードには離れた要素の出現を捉えられる利点があることによる．順序マイニングの手法（エピソードマイニング）により，エピソードは高速に抽出できることが知られている [3]．しかしエピソードは要素間にギャップを許すため，複数の素性に展開しなければならない問題がある．

この問題を解決するため，エピソードによる系列の書き換え手法を提案する．提案手法では 1 つのエピソード出現を 1 つの素性出現で扱える．このため，離れた要素の出現を考慮しつつ，素性数の増加を抑えられる．

2 章では扱う問題と関連研究について述べる．3 章では提案する系列の書き換え手法を述べる．4 章では実験と考察を行う．5 章で結論を示す．

2. 扱う問題と関連研究

2.1 扱う問題

本研究では状態を持つ系列の分類を行う．系列は順序を持った記号の並び $W = \langle w_1, w_2, \ldots, w_T \rangle$ とする．各記号 w_t は対応する状態 s_t を持つとする．状態 s_t の並び $S = \langle s_1, s_2, \ldots, s_T \rangle$ を状態列と呼ぶ．本研究ではクラス毎に状態遷移モデル $M = \{m_{c_1}, m_{c_2}, \ldots, m_{c_n}\}$ を構築する．各モデルとは，与えられた系列 W と状態列 S から W が S を持つ確率 $P(S|W)$ で定義される分布関数の生成モデルを示す．系列の記号および対応する状態の遷移に関する同時生成確率を尤度という．この最も高い値を与えるモデルを

[1] 法政大学理工学部創生科学科
Department of Advanced Sciences, Hosei University
[2] 島根大学大学院総合理工学研究科 情報システム学領域
Interdisciplinary Graduate School of Science and Engineering, Shimane University
[a)] jin.goto.9a@stu.hosei.ac.jp
[b)] shirai@cis.shimane-u.ac.jp

選択・分類する手法を最尤原理という．

$$c^* = \underset{c_k \in C}{\operatorname{argmax}} P(S|W)_{m_{c_k}} \quad (1)$$

本研究で扱う系列分類は，系列と状態列の対応を評価するため $P(S|W)$ の計算を必要とする．しかし，あくまで系列に単一のクラスを割り当てるものであり，系列各値の状態推定（系列ラベリング）ではない．

$P(S|W)$ を求めるために，状態に対応する系列値と状態遷移を扱えるモデルを用いる．主に隠れマルコフモデル（HMM）や最大エントロピーマルコフモデル（MEMM），条件付き確率場（CRF）の 3 つがある．本研究では，高い表現力を持つ CRF を分類に用いる．

CRF など最大エントロピーモデルの構築では，素性選択が大きな影響を与える．人手による素性記述は分野に依存した経験を必要とし，アノテータの個人差が生じる．系列は順序を持つため，順序の差異を反映できる素性が望ましい．ここで，本研究ではエピソードと呼ばれる頻出部分列に着目する．エピソードは，系列中にある窓幅内で出現するイベントの部分系列である．エピソードは N グラムと異なり一定の幅（窓幅）内であれば要素間にギャップを許す．このため離れた要素も考慮でき，N グラムより表現力が高い利点がある．順序マイニングの手法（エピソードマイニング）により，エピソードは高速に抽出できることが知られている [3].

エピソードマイニングの実行例を示す．**表 1** は順序マイニング用データの例である．$d1$ における要素数 2 のエピソードが各ウィンドウ内で出現している様子を示している．順序マイニングのしきい値は最小頻度 2，ウィンドウ幅 3，要素数 2 とする．この条件に従って抽出される頻出エピソードを**表 2** に示す．系列の部分列であるエピソード

表 1 マイニング対象の文書集合

t	1	2	3	4	5	6	7	8
d1	I	go	to	Tokyo	Disney	Land	.	
d2	She	went	to	Tokyo	Summer	Land	yesterday	.
d3	I	visited	Tokyo	yesterday	.			
d4	He	arrived	Tokyo	yesterday	.			
ウィンドウ	I|go, I|to		Tokyo|Disney, Tokyo|Land					
		go|to, go|Tokyo		Disney|Land, Disney|.				
			to|Tokyo, to|Disney		Land|.			

表 2 抽出される頻出エピソード（最小頻度 2，ウィンドウ幅 3，要素数 2）

エピソード	頻度
to|Tokyo	2
Tokyo|Land	2
Tokyo|yesterday	2
Tokyo|.	2
Land|.	2
yesterday|.	3

$\langle Tokyo, Land \rangle$ を Tokyo|Land のように表記する．N グラムでは間に異なる要素が入ったものは全て区別される．エピソードでは Tokyo|Land のように間に異なる語が出現しても同一のエピソード出現とみなすケースがある．例では Tokyo Disney Land と Tokyo Summer Land の 2 つの表現を Tokyo|Land でまとめている．

2.2 関連研究

マルコフモデルは，マルコフ性を持つ状態遷移からの記号出力により系列を生成するモデルである．このうち HMM は状態が未知である系列（不完全データ）からマルコフモデルを推定する．HMM は単純であるゆえ，モデル推定を高速に行える．しかし，HMM の素性には状態と記号のペアしか用いることができない．例えば，語数や，接尾語・接頭語等の機能語を対応させることができない．このように，HMM は重複する素性が扱えず，多様な素性表現が行えない．

MEMM は系列と状態列の対応を扱うマルコフモデルである．素性関数によって文字種や接頭辞などの部分的な情報も扱うことができる．しかしながら最適状態列の推定に際し，分岐の少ない経路に偏るラベルバイアス問題がある [2]．これは状態毎に最大エントロピーモデルの確率分布を構築しているためである．したがって状態列全体のもっともらしさを考慮できない．

条件付き確率場 (以下 CRF) は任意の構造を持つグラフ上のラベリングに適用できる識別モデルである．ラベリング対象が系列であるような CRF は Linear-Chain CRF(以下 LCCRF) と呼ばれ，HMM の拡張とみなせる [4]．LCCRF は以下のように $P(S|W)$ を直接求める．

$$P(S|W) = \frac{L}{Z(W)} \quad (2)$$

$$L = \exp(L1 + L2) \quad (3)$$

$$L1 = \sum_{t,i} \lambda_i f_i(W, s_t, s_{t-1}) \quad (4)$$

$$L2 = \sum_{t,i} \mu_i g_i(W, s_t) \quad (5)$$

$$Z(W) = \sum_{s_1, s_2, \ldots, s_T} \exp(L1 + L2) \quad (6)$$

$$(7)$$

ここで，f_i は入力系列 W, 現在の状態 s_t, 1 つ前の状態 s_{t-1} の 3 つ組に対し 0 または 1 を取る素性関数であり，以後遷移素性と呼ぶ．g_i は入力系列 W，現在の状態 s_t の組に対し 0 または 1 を取る素性で，以後状態素性と呼ぶ．λ_i と μ_i はそれぞれ f_i と g_i に対応する素性の重みである．遷移素性の重みはある状態からある状態への遷移しやすさを考慮している．状態素性の重みは現在の状態 s_t と周辺に出現した記号の対応を重みで表現している．L1 は W に対する遷移素性のスコア，L2 は状態素性のスコアを表している．$Z(W)$ は可能な状態列の全スコアの総和で，正規化項

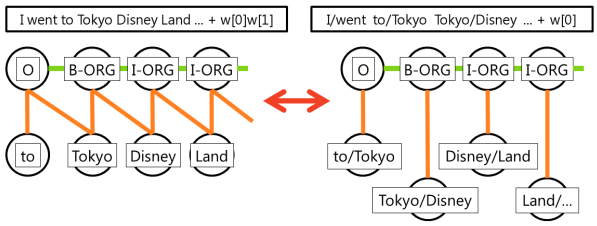

図 1 N グラム系列への書き換え例 (N=2)

である．

CRF は入力系列に対し大域的な確率分布を持つ点が MEMM と異なる．Lafferty ら [2] により，CRF は MEMM のようなラベルバイアス問題が起きないことが示されている．これらの特徴から，HMM より高い表現力を持ちつつラベルバイアスを起こさない LCCRF は状態を考慮した系列分類に最適である．したがって本研究では LCCRF モデルを用いて系列分類を行う．最大エントロピーモデルでは素性の表現が柔軟である分，識別に寄与する素性選択が重要となる．

3. 提案手法

本研究で提案する手法は 1 つのエピソード出現を 1 つの素性で扱うことを目的とする．この手法はユニグラム系列にバイグラム素性を適用したモデルと，バイグラム系列にユニグラム素性を適用したモデルが等価であることに基づく．後方 N-1 個のユニグラムと結合することにより，ユニグラムから N グラムへの書き換えは一意に行える．この書き換え操作により，ユニグラム素性を指定するだけで済む利点がある．提案手法では，系列の各要素（ユニグラム）をエピソードへ書き換える．エピソードは複数出現しうるため，書き換え先を 1 つに定める必要がある．本研究では書き換え対象の要素が状態を持つことを利用する．該当する状態と結びつきの強いエピソードへ書き換える．状態とエピソードの結びつきの強さは確信度として定義する．確信度は順序マイニング用データでの相対頻度で求める．系列の始端から終端まで同様の書き換え動作を繰り返すことで，ユニグラム系列全体をエピソード系列へ書き換える．

3.1 ユニグラムから N グラムへの書き換え

エピソードに類似する N グラムでは，ユニグラムの系列を N グラムの系列に一意に書き換えることができる．これは**図 1** のように表される．ここで，N グラムに書き換えた系列の各記号（N グラム）が 1 つの記号であるとみなしてユニグラム素性を適用する．このときユニグラム系列に N グラム素性を適用した場合と等価なモデルが得られる．図 1 の左側がユニグラム系列にバイグラム素性を適用する通常のモデルである．右側は，書き換えを行ってバイグラム系列にユニグラム素性を適用するモデルを示している．図 1 の B-ORG 状態の推定にはどちらも Tokyo, Disney の 2 語を用いている．同様に他の状態においても左右で同じ語を考慮している．すなわち両モデルは等価な状態素性を持つ．遷移素性も共通である．たとえ直接ユニグラム系列に N グラムの素性を対応付けることができなくても，書き換えにより同じモデルが得られる．この書き換え操作を，N グラムに類似したエピソードで行うのが提案手法である．指定する素性がユニグラム素性のみで済むことから，1 つのエピソード出現を 1 つの素性で表現できる．

3.2 エピソードの確信度計算

系列中の各ユニグラムをエピソードへ書き換える．書き換え先エピソードは，以下に定義する確信度が最も高いものとする．確信度は，書き換え対象記号 w_t，記号 w_t が持つ状態 s_t，現時刻で出現しており記号 w_t を含むエピソード e_k の 3 つ組から定まる．確信度 $conf$ を以下のように定義する．

$$conf(e_k, w_t, s_t) = \frac{freq(e_k, w_t, s_t)}{freq(e_k)} \tag{8}$$

確信度の計算例を示す．書き換え元の系列と，書き換え候補の頻出エピソードが既に得られているものとする．書き換え候補となる頻出エピソードは Tokyo|. と Tokyo|Land の 2 つとする．ここで，系列は単語列（文章），系列の各値が持つ状態は固有表現タグとする．B-ORG タグは組織名 (ORGANIZATION) の開始を示す．B-LOC タグは地名 (LOCATION) の開始を示す．

表 3 中，時刻 $t = 4$ の語 Tokyo を書き換えるとする．順序マイニング用データ**表 4** 中における各エピソードの頻度を**表 5** に示す．Tokyo|Land が 2 回出現したうち，語 Tokyo が B-ORG タグを持っていたケースは 2 回である．Tokyo|. が 3 回出現したうち，語 Tokyo が B-ORG タグを持っていたケースは 1 回である．書き換え対象の Tokyo は B-ORG タグを持つため，確信度 $conf(e_k, Tokyo, B-ORG)$ を求める．$conf(Tokyo|Land, Tokyo, B-ORG) = 2/2 = 1.0$, $conf(Tokyo|., Tokyo, B-ORG) = 1/3 = 0.33$ と計算できる．

3.3 ユニグラムからエピソードへの書き換え

確信度を求めたのち，最も高い確信度を持つエピソードを選択する．書き換え対象（Tokyo）が B-ORG タグを持っているため，B-ORG との確信度が最も高い Tokyo|Land を選択する．書き換え対象（Tokyo）をエピソード（Tokyo|Land）に置き換える．この操作を系列の始端から終端まで各値について行う．文末のピリオドなど，該当する箇所で出現するエピソードが 1 つも存在しない場合，要素をそのまま残す．この結果ユニグラム系列〈I, went, to, Tokyo, Disney, Land, yesterday, .〉は〈I, went, to, Tokyo|Land, Disney, Land|., yesterday|., .〉のように書き換えられる．

本手法では書き換え対象の要素の状態に応じて異なる

エピソードが選ばれうる．例えば，Tokyo の持つ状態が B-LOC であった場合，確信度が異なるため Tokyo|. へ書き換えられる．Tokyo1 語では地名 (LOC) であるか組織名 (ORG) であるかが曖昧である．しかし組織名タグを持つとわかっていれば，Tokyo|Land へ書き換えることで組織名であることがより明確になる．このような理由から，提案手法では系列の各値が状態を持つことを要求する．

表 3 書き換え元となる系列および状態列の例

t	1	2	3	4	5	6	7	8
S	O	O	O	B-ORG	I-ORG	I-ORG	O	O
W	I	went	to	Tokyo	Disney	Land	yesterday	.

表 4 順序マイニング用データ例

出現位置	t-1	t	t+1	t+2	t+3	t+4	t+5	t+6
d1	…	**B-ORG**	**I-ORG**	**I-ORG**	**O**			
	…	**Tokyo**	**Disney**	**Land**	**.**			
d2	…	**B-ORG**	**I-ORG**	**I-ORG**	**O**	**O**	O	O
	…	**Tokyo**	**Disney**	**Land**	**with**	**my**	family	.
d3	…	**B-LOC**	**O**					
	…	**Tokyo**	**.**					
d4	…	**B-LOC**	**O**					
	…	**Tokyo**	**.**					
		ウィンドウ (幅 5)						

表 5 確信度計算の例

	エピソード	
	Tokyo|Land	Tokyo|.
freq(e)	2	3
freq(e,Tokyo,B-ORG)	2	1
freq(e,Tokyo,B-LOC)	0	2
conf(e,Tokyo,B-ORG)	2/2=1.0	1/3=0.33
conf(e,Tokyo,B-LOC)	0/2=0.0	2/3=0.67

例で示した系列書き換え手続き（提案手法）を Algorithm 1 に示す．

4. 実験

4.1 実験の目的と手順

文書分類を素性セットが異なる LCCRF モデルで行い，比較する．ベースライン手法はユニグラム系列にバイグラム素性を適用する通常のバイグラムモデルとする．提案手法は系列をエピソードで書き換え，ユニグラム素性を適用するモデルとする．また，本研究における系列分類では系列の各値の状態推定（系列ラベリング）は行わないが，テストデータの系列の各値の状態推定も行って比較する．これは各 LCCRF モデルが状態遷移を正しくモデル化できているかを確かめるためである．本実験ではニュースコーパスデータを用いて順序マイニングを行う．この結果，頻出エピソードで学習用データおよびテストデータを書き換える．系列の各値は単語とし，各値が持つ状態は固有表現タグとする．

4.2 実験準備

Reuters Corpus Volume 1(RCV1) に収録されているニュース記事のうち，1996 年 10 月 1 日から 1996 年 12 月 31 日までの記事を実験に使用する．2 ヶ月分の記事を順序マイニング用データ，学習用データ，評価用テストデータの 3 つに分割する．分類先となるクラスは RCV1 のトピックコードに基づく．上位階層の CCAT(産業), ECAT(経済), GCAT(政治), MCAT(証券・株式市場) の 4 クラスとする．各記事のシングルラベル分類を行う．複数の大分類ラベルを持つ記事は除外する．また，固有表現が現れない記事も除外する．固有表現タグは RCV1 に付与されていないため，Stanford Named Entity Recognizer(CoNLL2003 4class IOB のモデル) を使用して正解タグを付与する．以下**表 6** に実験に用いるデータの詳細を示す．

表 6 実験に用いる記事件数

種別	マイニング用データ	学習用データ	テスト用データ
期間	10 月と 12 月	11 月 1 日から 11 月 14 日	11 月 15 日から 11 月 21 日
CCAT	28451	7538	3947
ECAT	6463	1458	904
GCAT	5486	1354	667
MCAT	18972	4186	2321
合計	59372	14536	7839

順序マイニングのしきい値は，最小頻度が各クラス中の文書数の 1.5%, 窓幅が 6, 要素数は 2 とする．また，ベースライン（N グラム素性モデル）は，$N = 2$ とし，要素数を同じ 2 とする．LCCRF モデルの学習においては疎な素性セットを用いる．したがって学習データ中に出現しない記号と状態の組み合わせは考慮しない．ベースライン手法も，素性数が同程度となるように（出現頻度に基づいて）素性集合を刈り込む．

実験の評価は系列分類結果と，状態推定結果に分けて行う．系列分類結果の評価は再現率と精度の調和平均である F 値，および正解率で行う．状態推定結果の評価は各クラスにおける再現率，精度，F 値それぞれのマクロ平均を用いる．

4.3 実験結果

表 7 に各モデルの学習時間と素性数を示す．学習時間は各モデルでほぼ同一となっている．素性数はクラス毎にばらつきはあるものの，GCAT クラスと MCAT クラスにおいて提案手法は両ベースライン手法の素性数の間になっている．CCAT クラスと ECAT クラスでは，提案手法の素性数が一番多くなっているが，ベースライン手法との素性数の差は 2%未満である．

表 8 に系列分類の結果を示す．最も高い F 値を与えて

表 10 状態推定結果の比較 (提案手法 / バイグラム素性モデル (頻度 6 以上))

	状態	精度	再現率	F1
CCAT	O	0.971 / 0.933 (+3.8%)	0.998 / 0.999 (-0.1%)	0.984 / 0.965 (+1.9%)
	I-PER	0.900 / 0.953 (-5.3%)	0.488 / 0.064 (+42.4%)	0.633 / 0.120 (+51.3%)
	I-ORG	0.932 / 0.935 (-0.3%)	0.853 / 0.523 (+33.0%)	0.891 / 0.671 (+22.0%)
	I-MISC	0.771 / 0.867 (-9.6%)	0.339 / 0.064 (+27.5%)	0.471 / 0.120 (+35.1%)
	I-LOC	0.894 / 0.967 (-7.3%)	0.717 / 0.386 (+33.1%)	0.796 / 0.552 (+24.4%)
	B-PER	0.909 / 0.960 (-5.1%)	0.312 / 0.039 (+27.3%)	0.465 / 0.074 (+39.1%)
	B-ORG	0.905 / 0.818 (+8.7%)	0.640 / 0.298 (+34.2%)	0.750 / 0.437 (+31.3%)
	B-MISC	0.960 / 0.972 (-1.2%)	0.736 / 0.121 (+61.5%)	0.833 / 0.215 (+61.8%)
	B-LOC	0.922 / 0.955 (-3.3%)	0.729 / 0.396 (+33.3%)	0.814 / 0.559 (+25.5%)
	マクロ平均	0.907 / 0.929 (-2.2%)	0.646 / 0.321 (+32.5%)	0.737 / 0.413 (+32.5%)
ECAT	O	0.960 / 0.935 (+2.4%)	0.998 / 0.999 (-0.1%)	0.978 / 0.966 (+1.3%)
	I-PER	0.857 / 0.800 (+5.7%)	0.248 / 0.114 (+13.4%)	0.385 / 0.200 (+18.5%)
	I-ORG	0.928 / 0.915 (+1.3%)	0.718 / 0.452 (+26.5%)	0.809 / 0.605 (+20.4%)
	I-MISC	0.899 / 0.963 (-6.4%)	0.378 / 0.163 (+21.5%)	0.532 / 0.279 (+25.4%)
	I-LOC	0.921 / 0.988 (-6.7%)	0.505 / 0.160 (+34.5%)	0.652 / 0.275 (+37.7%)
	B-PER	0.868 / 0.804 (+6.4%)	0.140 / 0.059 (+8.1%)	0.241 / 0.110 (+13.0%)
	B-ORG	0.934 / 0.943 (-0.9%)	0.656 / 0.448 (+20.8%)	0.770 / 0.607 (+16.3%)
	B-MISC	0.957 / 0.992 (-3.6%)	0.540 / 0.059 (+48.2%)	0.690 / 0.111 (+58.0%)
	B-LOC	0.946 / 0.966 (-2.0%)	0.577 / 0.297 (+28.0%)	0.717 / 0.454 (+26.3%)
	マクロ平均	0.919 / 0.923 (-0.4%)	0.529 / 0.306 (+22.3%)	0.642 / 0.401 (+24.1%)
GCAT	O	0.964 / 0.908 (+5.6%)	0.997 / 0.999 (-0.2%)	0.980 / 0.951 (+2.9%)
	I-PER	0.912 / 0.885 (+2.7%)	0.690 / 0.201 (+48.9%)	0.785 / 0.327 (+45.8%)
	I-ORG	0.859 / 0.892 (-3.3%)	0.732 / 0.282 (+45.0%)	0.790 / 0.428 (+36.2%)
	I-MISC	0.920 / 0.977 (-5.7%)	0.644 / 0.281 (+36.3%)	0.758 / 0.436 (+32.2%)
	I-LOC	0.921 / 0.978 (-5.7%)	0.742 / 0.395 (+34.7%)	0.822 / 0.562 (+26.0%)
	B-PER	0.948 / 0.977 (-2.9%)	0.576 / 0.166 (+41.0%)	0.717 / 0.284 (+43.3%)
	B-ORG	0.914 / 0.913 (+0.2%)	0.601 / 0.227 (+37.4%)	0.725 / 0.363 (+36.2%)
	B-MISC	0.955 / 0.970 (-1.5%)	0.800 / 0.236 (+56.4%)	0.871 / 0.380 (+49.1%)
	B-LOC	0.954 / 0.973 (-1.9%)	0.811 / 0.423 (+38.8%)	0.877 / 0.590 (+28.7%)
	マクロ平均	0.927 / 0.941 (-1.4%)	0.733 / 0.357 (+37.6%)	0.814 / 0.480 (+33.4%)
MCAT	O	0.974 / 0.942 (+3.3%)	0.998 / 1.000 (-0.2%)	0.986 / 0.970 (+1.6%)
	I-PER	0.930 / 0.956 (-2.6%)	0.426 / 0.200 (+22.6%)	0.584 / 0.331 (+25.3%)
	I-ORG	0.930 / 0.980 (-5.0%)	0.769 / 0.429 (+34.0%)	0.842 / 0.597 (+24.5%)
	I-MISC	0.849 / 0.943 (-9.4%)	0.673 / 0.333 (+34.0%)	0.751 / 0.492 (+25.9%)
	I-LOC	0.897 / 0.952 (-5.5%)	0.790 / 0.394 (+39.6%)	0.840 / 0.558 (+28.3%)
	B-PER	0.937 / 0.976 (-3.8%)	0.309 / 0.140 (+16.9%)	0.464 / 0.245 (+22.0%)
	B-ORG	0.917 / 0.949 (-3.2%)	0.593 / 0.276 (+31.7%)	0.721 / 0.428 (+29.3%)
	B-MISC	0.956 / 0.977 (-2.1%)	0.821 / 0.265 (+55.6%)	0.883 / 0.417 (+46.6%)
	B-LOC	0.932 / 0.964 (-3.2%)	0.775 / 0.357 (+41.8%)	0.847 / 0.522 (+32.5%)
	マクロ平均	0.925 / 0.960 (-3.5%)	0.684 / 0.377 (+30.7%)	0.769 / 0.506 (+26.2%)

Algorithm 1 ユニグラム系列からエピソード系列への書き換え手続き

入力: 長さ T のユニグラム系列 $W = \langle w_1, w_2, \ldots, w_T \rangle$,
長さ T の状態系列 $S = \langle s_1, s_2, \ldots, s_T \rangle$,
頻出エピソード集合 $E = \{e \mid freq_{min} \leq freq(e)\}$;
出力: 長さ T のエピソード系列 W^*;
$W^* = \langle\rangle$;
for $t = 1$ to T **do**
  $E_{appear} \leftarrow \{e_1, e_2, \ldots, e_n\}$;
  if $n \neq 0$ **then**
    $e^* \leftarrow \underset{e_k \in E_{appear}}{\text{argmax}} \; conf(e_k, w_t, s_t)$;
  else
    $e^* \leftarrow w_t$;
  end if
  W^* に書き換え後の記号 e^* を追加する
end for
return W^*;

表 7 学習時間および素性数

クラス		提案手法 (episode)	バイグラム (頻度 4 以上)	バイグラム (頻度 6 以上)
GCAT	学習時間 (秒)	3414	3253	3263
	タイプ数	89230	136774	72475
	素性数	106536	138834	73445
MCAT	学習時間 (秒)	969	1224	1152
	タイプ数	57646	76886	42518
	素性数	63114	77501	42845
CCAT	学習時間 (秒)	2328	2882	2484
	タイプ数	98677	111153	58857
	素性数	113898	112402	59365
ECAT	学習時間 (秒)	351	382	353
	タイプ数	32291	34847	19855
	素性数	35487	35224	20073

表 8 系列分類における再現率，精度，F 値および正解率

クラス		提案手法 (episode)	バイグラム (頻度 4 以上)	バイグラム (頻度 6 以上)
GCAT	再現率	1.000	0.952	0.939
	精度	0.867	0.667	0.660
	F1	0.929	0.784	0.775
MCAT	再現率	0.968	0.894	0.877
	精度	0.973	0.923	0.915
	F1	0.970	0.908	0.895
CCAT	再現率	0.983	0.929	0.928
	精度	0.958	0.902	0.892
	F1	0.970	0.915	0.910
ECAT	再現率	0.773	0.575	0.561
	精度	0.984	0.899	0.900
	F1	0.866	0.702	0.691
	正解率	0.956	0.879	0.871

表 9 バイグラム素性 (頻度 4 以上) のモデルによる状態推定結果

	状態	精度	再現率	F1
CCAT	O	0.933	0.999	0.965
	I-PER	0.932	0.066	0.122
	I-ORG	0.939	0.544	0.689
	I-MISC	0.872	0.062	0.116
	I-LOC	0.969	0.390	0.557
	B-PER	0.934	0.039	0.075
	B-ORG	0.836	0.309	0.451
	B-MISC	0.972	0.127	0.225
	B-LOC	0.956	0.391	0.556
	マクロ平均	0.927	0.325	0.417
ECAT	O	0.935	0.999	0.966
	I-PER	0.864	0.111	0.197
	I-ORG	0.924	0.459	0.614
	I-MISC	0.963	0.163	0.279
	I-LOC	0.988	0.160	0.275
	B-PER	0.879	0.057	0.108
	B-ORG	0.949	0.451	0.611
	B-MISC	0.993	0.062	0.117
	B-LOC	0.967	0.303	0.461
	マクロ平均	0.940	0.307	0.403
GCAT	O	0.909	0.999	0.952
	I-PER	0.896	0.206	0.334
	I-ORG	0.894	0.316	0.467
	I-MISC	0.969	0.293	0.450
	I-LOC	0.975	0.401	0.568
	B-PER	0.978	0.168	0.287
	B-ORG	0.917	0.231	0.369
	B-MISC	0.969	0.236	0.379
	B-LOC	0.973	0.423	0.590
	マクロ平均	0.942	0.364	0.489
x	O	0.942	1.000	0.970
	I-PER	0.954	0.195	0.324
	I-ORG	0.980	0.447	0.614
	I-MISC	0.941	0.340	0.499
	I-LOC	0.955	0.394	0.558
MCAT	B-PER	0.975	0.136	0.239
	B-ORG	0.954	0.281	0.435
	B-MISC	0.978	0.266	0.419
	B-LOC	0.965	0.357	0.521
	マクロ平均	0.960	0.380	0.509

いる CCAT クラスでは，ベースライン手法（頻度 4 以上）で F 値 0.915 に対し，提案手法では 0.970 と 5.5%上回っている．頻度 6 以上のベースライン手法（F 値 0.910）に対しては 6%の向上が見られる．最も低い F 値を与えている ECAT クラスにおいても，ベースライン手法（頻度 4 以上）で F 値 0.702, ベースライン手法（頻度 6 以上）で F 値 0.691 に対し，提案手法では F 値 0.866 と 16%以上高い．正解率では，提案手法が 0.956，ベースライン手法（頻度 4 以上）が 0.879 と 7.7%の向上が見られる．ベースライン手法（頻度 6 以上）の各指標は頻度 4 以上のものに比べ 0.1%から 2%ほど低く，準じた結果となっている．

表 9, **表 10** に各モデルによる状態推定の結果を示す．表 10 ではベースライン手法（頻度 6 以上のバイグラム素性）と提案手法を比較している．表 10 より，提案手法は F 値のマクロ平均で 24.1%(ECAT) から 33.4%(GCAT) の向上を示している．

4.4 考察

表 8 における F 値および正解率の向上から，提案手法は系列分類をうまく行えている．GCAT クラスではエピソード素性モデルのほうが素性数が 32298 個少ない (23%減少) にも関わらず，F 値は 14%以上高い．他のクラスでも，より少ない素性数あるいは同程度（2%未満の差）の素性数でよりよい分類結果を示している．したがって，エピソード素性はバイグラム素性に比べ少ない素性数でより高水準な分類が行えている．

表 10 において F 値が全てのクラスで 20%以上向上しており，提案手法はより正確に状態推定を行えている．また，推定対象のタグおよび精度，再現率で大きな差が見られる．表 10 の，CCAT クラスにおける O タグの推定結果を比較する．O タグの推定において，提案手法とベースライン手法は共に精度 93%以上，再現率 99%以上，F 値 96%以上を示している．一方，I-PER タグでは両手法で精度が 90%以上であるのに対し，再現率は提案手法で 48.8%, ベースライン手法で 6.4%と低い．他の状態およびクラスにおいても同様の傾向が見受けられる．これは，固有表現タグを持つ語に O タグ（非固有表現）を付与する誤りが多いためと考えられる．再現率の差から，この傾向はベースライン手法において特に顕著である．O タグ以外の再現率が低い原因として，未知記号の出現が考えられる．実際の状態推定の例から示す．

表 13 の 1 段目の事例では地名 Hong Kong のタグ付けに差が見られる．単語 Hong は両手法で同じ Hong|Kong へ書き換えられており，正しく B-LOC タグを付与している．単語 Kong では，提案手法は単語 Kong が残っており，ベースライン手法は直後の is とのバイグラム Kong|is になっている．ベースライン手法ではバイグラム Kong|is は未知記号として扱われており，誤って O タグを付与している．記号が未知であった場合，その状態は遷移素性のみで推定される．

ECAT クラスのモデルが持つ遷移素性のうち，例に関連する部分を**表 11** に示す．同様に状態素性も**表 12** に示す．また，提案手法の LCCRF モデルにおける状態遷移図を**図 2** に示す．矢印は状態の遷移を表し，付随する数値は対応する遷移素性の重みを示している．

表 12 より，エピソード素性では Hong|Kong, Kong, Turkey, UK に対応する状態素性が存在している．さらに，各記号と状態の対応を見ると，例で出現した正解状態の重みが最も大きくなっている．Hong|Kong では B-LOC が，Kong では I-LOC が，UK では B-LOC が最も大きな重みを示している．一方バイグラム素性では Hong|Kong 以外が未知であるため，状態素性が存在していない．したがって Kong, Turkey, UK などの地名タグの推定には遷移素性のみを用いている．表 11 では，両手法で O から O への遷移が最も大きな重みとなっている（エピソード素性で 6.99, バイグラム素性で 6.86）．したがって，遷移素性のみによって状態推定を行う場合 O から O への遷移が最も優先される．

バイグラム素性モデルが Kong|is の状態を推定する場合に注目する．遷移素性のみを用いるならば，前の Hong|Kong が持つ B-LOC からは I-LOC への遷移が起こるはずである．これはバイグラム素性モデルにおいて B-LOC から O への遷移素性の重みが 3.01, B-LOC から I-LOC では 5.65 であることによる．実際には O と推定しているが，これは後続の状態遷移を考慮していることによる．

Kong|is を I-LOC と推定した場合，次の時刻の状態推定では I-LOC から I-LOC への遷移素性が最も大きな重みを持つ．例のように後続に未知記号が続き，遷移素性のみを用いるとすると，I-LOC が繰り返される．一方，Kong|is の状態を O とすると遷移素性に従って O が続く．I-LOC から I-LOC への状態遷移が 4.86 の重みを持ち，O から O への状態遷移は 6.86 の重みを持つ．従って系列長が長くなるほど，I-LOC が続く場合に比べ O が続く状態列のスコアが大きくなる．例の場合，時刻 2 の時点で O が続く状態列のスコアが I-LOC が続く状態列のスコアより高くなる．LCCRF モデルは，系列の各値および状態それぞれに対応する素性の重みを全ての時刻で足し合わせて評価する．したがって局所的には優先されない遷移でも，全体のスコアが高くなるものが選択される．O から O への遷移が続く状態列は全体のスコアが高くなりやすいため，O タグが付与されると考えられる．

表 13 に評価用テストデータにおける状態推定結果の一部を示す．ベースライン手法は頻度 6 以上のバイグラム素性のモデルである．エピソードおよびバイグラムのうち，下線が引かれたものは未知記号である．表 13 の 3 段目では，エピソード素性モデルも語 PSBR に O タグを誤って付

表 11 ECAT クラスのモデルにおける遷移素性の重み

s_{t-1}	s_t	重み (エピソード)	重み (バイグラム)
O	O	6.99	6.86
B-ORG	I-ORG	6.60	6.32
I-ORG	I-ORG	6.03	6.22
O	B-ORG	5.97	6.59
O	B-LOC	5.59	6.49
B-LOC	I-LOC	5.48	5.65
I-LOC	I-LOC	4.65	4.86
B-LOC	O	2.89	3.01
B-ORG	O	2.36	2.50
I-ORG	O	2.11	2.77
I-LOC	O	2.11	2.42
B-LOC	B-ORG	-0.87	-0.47

表 12 ECAT クラスのモデルにおける状態素性の重み

W	s_t	重み (エピソード)	重み (バイグラム)
Hong|Kong	B-ORG	0.87	2.21
Hong|Kong	I-ORG	0.52	1.21
Hong|Kong	B-LOC	3.14	4.48
Kong	I-ORG	1.63	-
Kong	I-LOC	4.28	-
Kong|is	I-LOC	-	-
Turkey	B-LOC	4.56	-
Turkey|may	B-LOC	-	-
UK	O	-2.72	-
UK	B-ORG	0.39	-
UK	I-ORG	-0.07	-
UK	B-LOC	3.61	-
UK	B-MISC	-0.01	-
UK|public	B-LOC	-	-
PSBR	B-ORG	3.57	-
PSBR	I-ORG	3.79	-
PSBR|-RRB-	B-ORG	-	-
-RRB-|on	O	1.56	-
-RRB-|on	I-ORG	-0.67	-

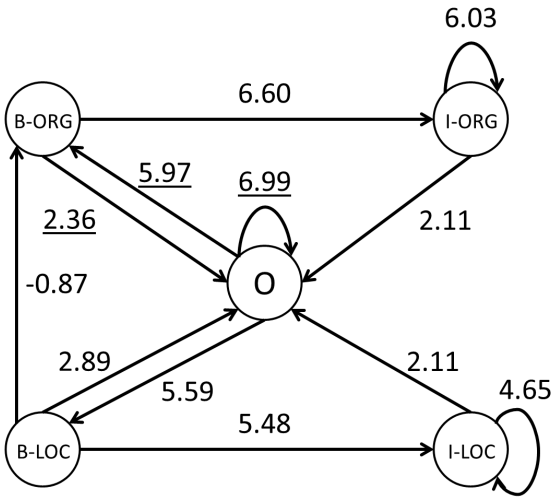

図 2 ECAT クラスにおける，提案手法モデルの状態遷移図 (O, ORG, LOC のみ)

与している．表 12 より，語 PSBR が B-ORG を持つ状態素性は存在しており，3.57 の重みを持っている．エピソード素性モデルは重みを正しく学習できており，原因は未知記号の出現ではない．この誤りは，学習データ中の状態遷移の傾向によるものと考えられる．

表 11 より，O から B-ORG への遷移素性の重みが 5.97, B-ORG から O への遷移素性の重みが 2.36 となっている．一方 O から O への遷移素性は 6.40 となっている．語 PSBR の前後〈-LRB-,PSBR,-RRB-〉の正解状態列は〈O,B-ORG,O〉である．〈O,B-ORG,O〉での，遷移素性によるスコアは表 11 より $5.97 + 2.36 = 8.33$ となる．状態素性も加味すると $8.33 + 3.57 = 11.9$ となる．〈O,O,O〉でも同様にスコアを求める．語 PSBR と状態 O の状態素性は存在しないため，遷移素性の重みの和がスコアとなる．O タグへの遷移が起きているから，$6.99 + 6.99 = 13.98$ となる．$13.98 > 11.9$ より，1 つ後ろまで考慮すると O タグを付与するほうがスコアが高くなることが分かる．このように，状態素性の重み以上に O から O への遷移の影響が強いため，誤った状態推定を行っている．各素性の重みの大きさは学習データ中における頻度が影響している．したがってこの誤りは提案した手法によるものではなく，学習データ中の状態遷移の傾向によるものと考えられる．

5. 結論

本研究では順序マイニングによって獲得した頻出エピソードを系列分類に用いる手法を提案した．確信度に基づく書き換えを行うことにより，エピソードの出現と素性の出現を対応付けた．エピソード素性を用いた系列分類実験では正解率 95.6%と，同程度の素性数のバイグラム素性の場合 (87.9%, 87.1%) に比べて正解率が約 7%向上した．同様のモデルを用いた状態推定においても提案手法はベースライン手法に比べて F 値が最大で 33%向上した．未知の記号出現がバイグラム素性に比べて生じにくい点を確かめ，状態推定がより正確に行えることを示した．今後の発展として，本研究では明示的に与えられると仮定した状態が隠れている，隠れ条件付き確率場モデルが考えられる．

謝辞 香港中文大学の Wai Lam 先生には隠れ条件付き確率場についての研究動向や条件付き確率場の代表的タスクおよびコーパスについて大変貴重な御意見と助言をいただきました．深く感謝いたします．

参考文献

[1] 最大エントロピーモデルのための素性選択の簡素化, 2016 年電子情報通信学会総合大会 ISS 特別企画「学生ポスターセッション」, 2016, 九州大学, 福岡共著 (後藤 仁, 三浦 孝夫), 平成 28 年 (2016) 3 月

[2] J. Lafferty, A. McCallum, F. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, ICML 2001

[3] Mannila, Heikki, Hannu Toivonen, and A. Inkeri Verkamo. "Discovery of frequent episodes in event sequences." Data mining and knowledge discovery 1.3 (1997): 259-289.
[4] Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields." Foundations and Trends in Machine Learning 4.4 (2012): 267-373.
[5] Uno, Takeaki, et al. "An efficient algorithm for enumerating closed patterns in transaction databases." International Conference on Discovery Science. Springer Berlin Heidelberg, 2004.
[6] Xing, Zhengzheng, Jian Pei, and Eamonn Keogh. "A brief survey on sequence classification." ACM Sigkdd Explorations Newsletter 12.1 (2010): 40-48.

表 13 テストデータ系列の各値の状態推定結果例（ECAT クラスより抜粋）

原文	Hong	Kong	is	among	the	world	's	most
エピソード系列	Hong|Kong	Kong	is|'s	among	the|'s	world	's	most
バイグラム系列	Hong|Kong	Kong|is	is|among	among|the	the|world	world|'s	's|most	most|efficient
正解状態列	B-LOC	I-LOC	O	O	O	O	O	O
推定状態列 (エピソード)	B-LOC	I-LOC	O	O	O	O	O	O
推定状態列 (バイグラム)	B-LOC	O	O	O	O	O	O	O
原文	Turkey	may	risk	a	crisis	next	year	with
エピソード系列	Turkey	may|a	risk	a|year	crisis	next	year|with	with|'s
バイグラム系列	Turkey|may	may|risk	risk|a	a|crisis	crisis|next	next|year	year|with	with|speculative
正解状態列	B-LOC	O	O	O	O	O	O	O
推定状態列 (エピソード)	B-LOC	O	O	O	O	O	O	O
推定状態列 (バイグラム)	O	O	O	O	O	O	O	O
原文	UK	public	sector	borrowing	requirement	-LRB-	PSBR	-RRB-
エピソード系列	UK	public|sector	sector	borrowing	requirement	-LRB-	PSBR	-RRB-|on
バイグラム系列	UK|public	public|sector	sector|borrowing	borrowing|requirement	requirement|-LRB-	-LRB-|PSBR	PSBR|-RRB-	-RRB-|data
正解状態列	B-LOC	O	O	O	O	O	B-ORG	O
推定状態列 (エピソード)	B-LOC	O	O	O	O	O	O	O
推定状態列 (バイグラム)	O	O	O	O	O	O	O	O