4B-05

スケルトン化を用いた周期パターン発見に関する一考察

大滝 啓介 †‡　　山本 章博 †
京都大学大学院 情報学研究科
ootaki@iip.ist.i.kyoto-u.ac.jp　akihiro@i.kyoto-u.ac.jp

1　はじめに

データマイニングにおける基本問題である**パターン発見問題**において，データベース中のレコードが時間情報を持つことは様々な問題における自然な仮定であり，パターンの例として系列パターンがある [1, 2]．**周期パターン発見問題**とは特に周期に着目して系列パターンを発見する問題であり，周期性は時系列データの基本的な特徴付けであるため，これまで広く研究されている [3]．しかしパターン発見によって列挙されるパターンは**冗長**であることが知られている．そのため列挙されたパターンから特徴的なパターンのみを選択する問題が研究されているが [4]，未だに困難な問題の一つである．本稿は [5] において，系列パターン発見における冗長性を解決するために提案された，**スケルトン化**を周期パターンに対して拡張し，冗長性の問題に関して議論する．

2　準備

周期パターン発見問題　アルファベット Σ 上の列 $S \in \Sigma^+$ と周期 p，パラメータ θ が与えられる．列 S の周期 p による $m = \lceil |S|/p \rceil$ 分割を $S = \langle ps_1, \ldots, ps_m \rangle$ と表す．周期パターン発見問題において完全な周期性を持つ列は非常に稀であるため，任意の文字を表す記号 $\star$ を導入し，$p \in (\Sigma \cup \{\star\})^+$ を**パターン**と呼ぶ．パターン p が ps_i に部分列として出現する場合，$p \preceq ps_i$ と表す．列 S における p の**頻度**を $\mathrm{Sup}_{\mathrm{P}}(p) = |\{ps_i \mid p \preceq ps_i\}|/|m|$ と定義し，$\mathrm{Sup}_{\mathrm{P}}(p) \geq \theta$ が成り立つすべてのパターンを発見する問題を，**頻出部分周期パターン発見問題**と呼ぶ．

系列パターン発見とスケルトン化　系列データベース $\mathrm{DB} = \{S_i \mid S_i \in \Sigma^+\}$ とパラメータ θ に対して，系列パターン $s \in \Sigma^+$ の頻度を $\mathrm{Sup}(s) \equiv |\{S_i \mid S_i \in \mathrm{DB}, s \preceq S_i\}|/|\mathrm{DB}|$ と定義し，$\mathrm{Sup}(s) \geq \theta$ を満たす全てのパターンを発見する問題が系列パターン発見問題である．

既存研究 [5] に提案された**スケルトン化**とは，DB に出現する文字 $c \in \Sigma$ の出現位置を観察することで，文字のクラスタを構築する前処理手法である．具体的には DB から以下のような**類似度グラフ**を構成する．

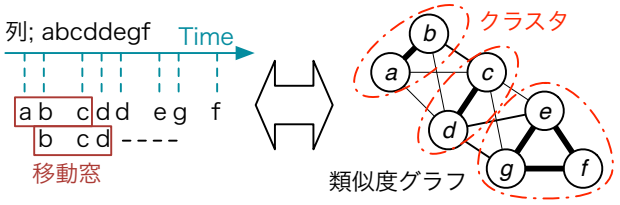


図 1: スケルトン化と類似度グラフ

定義 1 (類似度グラフ) 系列データベース DB から構築される類似度グラフ $G_{\mathrm{DB}} = (V, E)$ において，頂点 $v \in V$ はある文字 $c \in \Sigma$ に対応し，頂点 s, t 間の重み $W_{s,t}$ は

$$W_{s,t} = N^{-1} \sum_{1 \leq n \leq N, e_i, e_j \in S_n} \delta\left(|l(e_i, S_n) - l(e_j, S_n)| \leq r\right) \quad (1)$$

として求める．関数 $l(e_i, S_n)$ は列 S_n における文字 e_i の出現位置を返し，r は移動窓の幅を表す．

スケルトン化では，構築された類似度グラフ G_{DB} に対して**グラフからのクラスタ発見手法** (例えばスペクトラムクラスタリング [6]) を適用することで，DB における各文字の出現傾向に基づいたクラスタを発見する．

系列データベース DB から発見される系列パターン $S \in \Sigma^+$ は，前提より，何らかの時間的な前後関係を表している．そこで DB 中の多くの列において，**近くに出現する文字は，時系列上の何らかの関係を持つ**という暗黙の仮定が，式 1 に反映されている．

3　提案手法

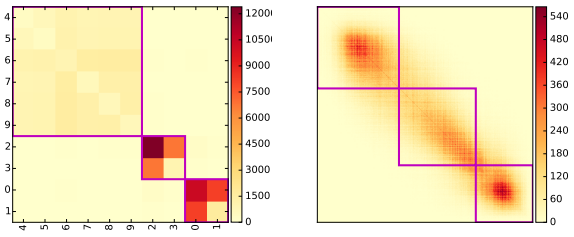
本稿ではスケルトン化の手法と議論を周期パターン発見に対して展開する．つまり**パターン発見問題における暗黙の前提をデータの前処理に積極的に適用する．**

周期スケルトン化　周期パターンにおける暗黙の前提とは列 S を m 個の部分列に分割し，パターン p の頻度 $\mathrm{Sup}_{\mathrm{P}}(p)$ を定めることである．そのため，式 1 に従って重み計算を行う際，幅 r の移動窓をパラメータ p により

Mining periodic patterns of sequences via skeletonization
Keisuke Otaki †‡　Akihiro Yamamoto †
†Graduate School of Informatics, Kyoto University
‡JSPS Research Fellow (DC2)

表 1: 利用したデータ

Data	Length	$\|\Sigma\|$	Period	# of clusters
SYNTH	600	10	3	3
KYOTO	43,833	359	365	unknown



(a) SYNTH　(b) KYOTO

図 2: 類似度グラフとクラスタ

周期拡張することで，周期性を考慮したスケルトン化を構築する．これを**周期スケルトン化**と呼ぶ．

形式的にはある列 S 上の位置 $i, j \in \{0, \ldots, |S|\}$ $(i < j)$ について，文字 $S[i]$ と $S[j]$ に対応するノードを s, t と置くと，重み $W_{s,t}$ を以下のように更新する．

$$W_{s,t} \leftarrow W_{s,t} + \begin{cases} d_1(i,j) & ||i-j|| \le r, \\ d_2(i,j) & j \equiv i \pmod{p}, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

ここで $d_1(\cdot,\cdot)$ と $d_2(\cdot,\cdot)$ はある位置 i の近傍と，その周期拡張された位置 $i+p, i+2p, \ldots$ について重みを計算するための関数であり，デルタ関数 $\delta(\cdot)$ やガウシアンカーネル $\exp(\cdot)$ によって実現される．以降では最も簡単なデルタ関数 $\delta(\cdot)$ を両方に用いる．

4　実験と考察

周期スケルトン化を用いて得られるクラスタにより入力列を符号化し，周期パターン発見を適用することで，周期スケルトン化の影響を確認・考察する．

実験データ (表 1)　周期 3 のサイクルを持つ HMM によって生成した人工データ SYNTH と，離散化した時系列 KYOTO を利用する．

類似度グラフとクラスタ　各データから得られた類似度グラフをヒートマップにより表現したものを図 2 に示す．それぞれのヒートマップ (図 2a,2b) において紫色の矩形は発見されたクラスタを表している．

周期パターン発見への応用　データ KYOTO を例にとり，周期性の前提から得られたクラスタが，周期パターン発見にどのように影響するかを調査する．求められた 3 つのクラスタをサイズの大きい方から順番に C_1, C_2, C_3 とする．長さ $|S| = 43,833$ の列 S 中の記号を C_1, C_2, C_3 の順にラベルで置き換え，列 S_1, S_2, S_3 を作る．これに周期パターン発見アルゴリズムを適用し，得られた頻出周期パターンの個数を表 2 に示す．

表 2: 得られた頻出周期パターン (パラメータ θ)

Data	$\|\Sigma\|$	$\theta = 0.9$	0.7	0.5
KYOTO	359	0	0	0
S_1	224	9,065	57,596	133,027
S_2	97	28,134	210,806	523,021
S_3	3	54,354	349,648	917,403

結果と考察　クラスタラベルを用いた再符号化により，高い閾値 θ に対しても頻出周期パターンを発見した．アルファベット Σ によって可能な全てのパターンの個数は一般に膨大であるから，通常のパターン発見手法を適用しても，これらのパターンは θ を小さく (例えば $\theta = 0.05$) するまで発見することが出来ない．このとき，そのような非情に稀なパターンはノイズであることを否定できず，評価することが難しい．

実験によりクラスタラベルを用いることで，通常のスケルトン化と同じく，周期パターンに関しても，データを全体的に特徴付けるような，大まかなパターンを列挙することが可能であるという利点を確認した．

5　まとめ

本稿では系列パターン発見におけるスケルトン化を，周期パターン発見に対して一般化させた上で，パターン発見問題が抱えている冗長性に関して議論した．今後の課題として，複数周期が重なっているデータに関する適用や，よりパターン発見手法と組み合わせたアルゴリズムの開発，理論的解析などがある．

参考文献

[1] R. Agrawal and R. Srikant. Mining Sequential Patterns. In *11th ICDE*, pp. 3–14, 1995.

[2] J. Pei, J. Han, B. Mortazavi-asl, H. Pinto, Q. Chen, U. Dayal, and M-C. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *17th ICDE*, pp. 215–224, 2001.

[3] J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. In *15th ICDE*, pp. 106–115, 1999.

[4] J. Yang, W. Wang, and P.S. Yu. Infominer: Mining surprising periodic patterns. In *7th KDD*, pp. 395–400, 2001.

[5] C. Liu, K. Zhang, H. Xiong, G. Jiang, and Q. Yang. Temporal skeletonization on sequential data: Patterns, categorization, and visualization. In *20th KDD*, pp. 1336–1345, 2014.

[6] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 13*, pp. 849–856, 2001.