3I4-1

区間イベント系列からの頻出部分グラフマイニング

Frequent Subgraph Mining in a Long Sequence of Interval-based Events

鈴木湧人　　尾崎知伸
Yuto Suzuki　　Tomonobu Ozaki

日本大学文理学部
College of Humanities and Sciences, Nihon University

In this paper, we propose a framework to obtain useful structures or patterns from multi-dimensional time series data. By regarding each time series as a node and by extracting similar subsequences between nodes as edges, we obtain a long sequence of interval-based event, i.e. edges with time-interval. Frequent subgraph patterns are enumerated from the obtained sequence by an extended subgraph mining algorithm. We confirm the effectiveness of the proposed framework by preliminary experiments on financial data.

1. はじめに

多次元時系列データからのパターンマイニング，すなわち特徴的な構造やパターンを抽出する技術に関する既存研究の多くは，各時系列データに繰り返し現れる部分系列を記号化し，その上で他の系列との関係性を抽出するという方法を採用している．これに対し本論文では，各系列の部分系列の類似性に着目することで，既存手法では捉えることのできないパターンの抽出を試みる．本論文で提案する手法を gIE^T(Time based Subgraph Pattern of Interval Events) と表記する．gIE^T は，長大な複数の区間系列データ（多次元時系列データ）を入力とし，辺にイベント時間を持つ頻出部分グラフ構造を出力する．具体的には，各系列を頂点，系列同士の部分類似区間を辺，辺ラベルを部分類似区間の開始時刻と終了時刻とする多重無向グラフを考え，そこからイベント時間（辺に付与された開始時間と終了時間）を考慮した部分グラフパターンを列挙する．

これまでに，区間イベント系列やグラフ，グラフ系列を対象とした頻出パターン発見手法が多数提案されている [1, 2, 3, 4]．TPrefixSpan [1] は，点系列の集合から最右拡張を用いて縦型にパターンを列挙するアルゴリズム PrefixSpan [5] を区間系列に拡張したアルゴリズムである．また gSpan[2] は，ラベル付きの無向グラフ集合から頻出する部分グラフを列挙する代表的なアルゴリズムである．一方，グラフ系列，すなわち単一グラフの時間的な変化を対象とした手法としては，GERM[3] や LF-Rules[4] があげられる．これらは gSpan の技術を元にグラフ系列から相関ルールを列挙する．GERM では，辺を（時間幅を持たない）イベントとして捉えて部分グラフを列挙する．これに対し提案手法 gIE^T は，辺を時間幅を持つ区間イベントとして捉える点が最大の違いである．

2. 区間イベント系列とグラフパターンの列挙

2.1 区間イベント系列

本論文では，多次元時系列データから，各系列を頂点，系列同士の部分類似区間を辺，辺ラベルを部分類似区間の開始時刻と終了時刻とする多重無向グラフ $G = (V, E)$ を獲得する．ここで頂点集合 V は系列の集合，辺集合 E は区間イベントの集合である．また，辺 $e = (v_1, v_2, t_s, t_e) \in E$ は区間イベントであり，２つの系列データ $v_1, v_2 \in V$ が，時刻 $[t_s, t_e]$ において類似であることを表す．なおグラフ G は，時間軸に沿った区間イベント（辺）系列と等価であると判断できる．図 1 に，区間イベント系列とグラフ G の例を示す．

$E =< e_1, e_2, e_3, e_4, e_5 >$
$e_1 = (A, B, 1, 5)$
$e_2 = (A, B, 3, 7)$
$e_3 = (A, B, 2, 6)$
$e_4 = (A, C, 1, 8)$
$e_5 = (B, D, 1, 5)$

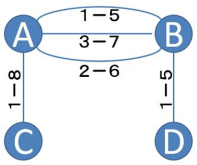

図 1: 区間イベント系列とそのグラフ表現

2.2 部分グラフの列挙

提案手法 gIE^T は，多重無向グラフ G と最小支持度 σ，最大時間差 ω を入力とし，G 中に σ 回以上出現する，辺の時間差が ω 以内の単純連結グラフ P を部分グラフパターンとして列挙する．なお，部分グラフパターンの支持度は，単一グラフに対する支持度 [6] を援用する．また，同型グラフの列挙を避けるため，gSpan や GERM 等で採用されている DFS Code と Lexicographic Order に基づく標準形（正準形）を，辺の開始・終了時間を考慮するよう拡張し利用する．

図 2 に提案アルゴリズム gIE^T を示す．図中において，G は入力グラフ，P は結果として得られるパターンの集合，g はグラフパターンをそれぞれ表す．gIE^T は，GERM[3] 同様，最右拡張を用いた深さ優先探索アルゴリズムであり，$isCanonical(g)$ は標準形判定，$RMO(g)$ は最右拡張，$support(g)$ は支持度計算，$timediff(g)$ は時間差の計算をそれぞれ表す．

3. 評価実験

提案手法の有効性を評価するため，Java 言語を用いて gIE^T を実装し評価実験を行った．また比較のため，GERM[3] のパターン列挙と本質的に同等の手法として，区間イベントの開始時間のみを考慮することで区間イベントを点イベントとして扱う手法（$GERM'$）の実装も行った．

連絡先: 尾崎 知伸，日本大学 文理学部 情報科学科，〒 156-8550
東京都世田谷区桜上水 3-25-40，tozaki@chs.nihon-u.ac.jp

$gIE^T(G, P, g)$

1: **if** $g \neq isCanonical(g)$ **then**
2: return
3: $P \leftarrow P \cup \{g\}$
4: **for each** $g' \in RMO(g)$
5: **if** $support(g') \geq \sigma \wedge timediff(g') \leq \omega$ **then**
6: $gIE^T(G, P, g')$

図 2: アルゴリズム gIE^T

実験には株価データ [*1] を使用した．2007 年から 2015 年の間での取引のあった約 2200 日分のデータから，無作為に 50 社を選択し，各会社の株価を一つの系列とした．これら 50 の系列に CrossMatch アルゴリズム [7] を適用することで類似部分系列の抽出し，結果として得られる区間イベント系列（総数 13,444，1 日毎の区間イベント数の平均 6.13，分散 23.27）を実験データとした．実験では，最小支持度 $\sigma \in \{10, 15, 20\}$ と最大時間差 $\omega \in \{7, 14, 21, 28\}$ を変化させ，実行時間とパターン数を集計した．結果を表 1 に示す．表中において，'–' は 1 時間以内に列挙が終わらなかった場合を表す．

実験結果より，gIE^T では最小支持度 σ が 20 以上のときは最大時間差 ω に関係なくパターンが抽出されないことが分かる．一方，$GERM'$ では最小支持度 σ が 20 以上のときは最大時間差 ω が $\{7, 14, 21\}$ のときにパターンが抽出されることが分かる．また，パターン数の変化に関して gIE^T と $GERM'$ を比較すると，両者ともに同様の傾向があることが読み取れる．

図 3 に，最小支持度 $\sigma = 10$，最大時間差 $\omega \in \{7, 14, 21, 28\}$ の場合に，gIE^T により得られるパターン数のサイズ（辺の数）毎の分布を示す．図より，サイズ 5 及び 6 のパターンが多い傾向があることが分かる．

表 1: 実験結果

	パターン数			実行時間 (秒)		
$\omega \backslash \sigma$	10	15	20	10	15	20
gIE^T						
7	22	3	0	21	0	0
14	3241	35	0	54	1	0
21	34984	66	0	2100	1380	0
28	36260	32260	0	3180	2760	0
$GERM'$						
7	206	88	29	3	1	0
14	25767	15254	6342	1018	781	961
21	–	–	13915	–	–	1050
28	–	–	–	–	–	–

4. まとめ

本論文では，多次元時系列データからグラフを構成し，そこからイベント時間を考慮したパターンを列挙する手法を提案した．今後の課題として，提案した手法の有用性を評価するため

*1 http://k-dbcom/

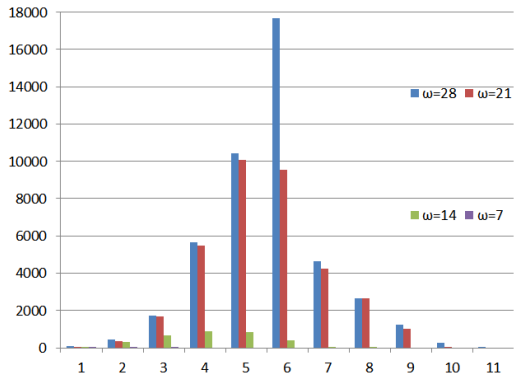

図 3: パターン数の分布（x 軸はパターンのサイズ）

に他のデータでの再実験と，類似区間だけではなく非類似区間に適用することが挙げられる．

参考文献

[1] S.-Y. Wu and Y.-L. Chen : Mining Nonambiguous Temporal Patterns for Interval-Based Events, *IEEE Transactions on Knowledge and Data Engineering*, Vol.19, No.6, pp.742-758 (2007).

[2] X. Yan and J. Han : gSpan: Graph-Based Substructure Pattern Mining, *Proc.* of the 2002 IEEE International Conference on Data Mining, pp.721-724 (2002).

[3] M. Berlingerio, F. Bonchi, B. Bringmann and A. Gionis : Mining graph evolution rules, *Proc.* of the 2009 European Conference on Machine Learning and Knowledge Discovery in Databases, pp.115-130 (2009).

[4] C. W.-k. Leung, E.-P. Lim, D. Lo and J. Weng : Mining interesting link formation rules in social networks, *Proc.* of the 19th ACM International Conference on Information and Knowledge Management, pp.209-218 (2001).

[5] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M. Hsu : Prexspan: Mining Sequential Patterns by Prex-Projected Pattern Growth, *Proc.* of the 17th International Conference on Data Engineering, pp.215-224 (2001).

[6] B. Bringmann and S. Nijssen : What Is Frequent in a Single Graph?, *Proc.* of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp.858–863 (2008)

[7] M. Toyoda, Y. Sakurai and Y. Ishikawa : Pattern discovery in data streams under the time warping distance, *The International Journal on Very Large Data Bases*, Vol.22, No.3, pp.295-318 (2012)