因果関係を考慮した連鎖パタンマイニング手法のパラメータの設定

因果関係を考慮した連鎖パタンマイニング手法のパラメータの設定

因果関係を考慮した連鎖パタンマイニング手法のパラメータの設定

○大久保勇輔　李セロン　岡田吉史（室蘭工業大学）

概要　我々は以前, 複数の系列データにまたがって繰り返される頻出パタン集合である連鎖パタンを抽出する方法を提案した．しかしながら，この方法では，同時刻帯に偶然出現している偽の連鎖パタンが抽出されてしまう問題が残されていたため，因果関係を考慮した連鎖パタン抽出法を開発した．本研究では，多種多様な人工データセットを用いた従来法との比較実験を通して，パラメータの設定方法について検証を行う．

キーワード: シーケンシャルパタンマイニング，連鎖パタンマイニング,因果関係，データマイニング

1　はじめに

近年，時々刻々と増加し続ける膨大な時系列データから，有用な情報や知識を発見する系列パタンマイニング技術が注目されている．我々はこれまで，Fig. 1 に示されるような複数の系列データにまたがって繰り返し現れるパタン群（以下，連鎖パタン）を発見する連鎖パタンマイニング手法（以下，前手法）を開発してきた[1) 2)]．

本研究では，前手法では考慮されなかった出現パタン間の因果関係に基づく，新しい連鎖パタンマイニング法を開発する．この方法は，前提として解析者が着目する系列（着目系列）とそこに出現するパタン（結果パタン）があると仮定し，その他の系列（原因系列）から，誘導原因となるパタン（原因パタン）を特定する．パタン間の因果関係の推定は、重み付き有向グラフにより実現される．本手法を用いることで，例えば，複数のバイタルデータから，興味の事象（結果パタン）の発生原因（原因系列や原因パタン）を見つけ出すことができるようになると期待される．

以下本稿では，人工的に生成したデータセットを用いて，本手法と前手法との抽出精度の比較を行い，パラメータの設定方法について検討した結果を報告する．

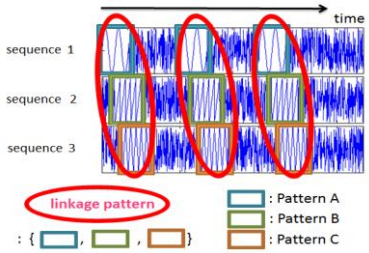

Fig.1: Example of linkage pattern

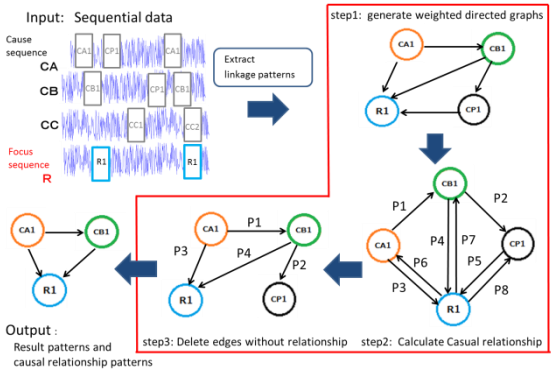

Fig.2: Procedure of this method

2　方法

Fig.2は本手法の流れである．まず，前手法により連鎖パタンの抽出を行う．本手法は，抽出された連鎖パタンに含まれるパタン間の因果関係を表現するために，因果の確率をエッジに付与した有向グラフ（重み付き有向グラフ）を用いることにより実現される．

以下，Fig.2を使って本手法の処理手順を説明する．ここでは，　系列1，系列2，系列3をそれぞれ原因系列CA, CB, CCとし，系列4を着目系列Rと呼ぶ．原因系列Cに出現するパタンを原因パタンCA_n, CB_n, CC_nとし，着目系列Rに出現するパタンを結果パタンR_xとする．ここで，n，x はパタンに付与される番号である．

2.1 重み付き有向グラフの作成

Fig.2 の step1 では，重み付き有向グラフを作成している．図のように同時刻帯に出現するパタンを時系列順に有向グラフで結び，2.2 節の方法に基づいて各パタン間の因果関係の数値をエッジに付与させる．

2.2 パタン間の依存関係の数値化

2.2.1 原因パタンと原因パタンの依存関係の数値化

Fig.2 の step2 を用いて原因パタン間における依存関係の数値化を説明する．CA_iの出現頻度を $occ(CA_i)$とし，各連鎖パタンが出現する時間帯においてCA_iの次にCB_jが出現する頻度を $occ(CA_i \rightarrow CB_j)$とする．このとき，$CA_i$から$CB_j$の依存関係の度合を $P(CA_i \rightarrow CB_j)$は下式により算出される：

$$P(CA_i \rightarrow CB_j) = occ(CA_i \rightarrow CB_j) / occ(CA_i) .$$

2.2.2 原因パタンと結果パタンの因果関係の数値化

次に，Fig.2 の step2 を用いて原因パタンと結果パタンの因果関係の数値化を説明する．R_kの出現頻度を $occ(R_k)$，各連鎖パタンが出現する時間帯においてCA_iが発生したとき，それより後にR_kが出現する頻度を $occ(CA_i \rightarrow R_k)$，$R_k$が発生したとき，$R_k$より前に$CA_i$が発生している頻度を $occ(R_k \rightarrow CA_i)$とする．このとき，因果関係の度合 $P(CA_i \rightarrow R_k)$，$P(R_k \rightarrow CA_i)$は下式により算出される：

$$P(CA_i \rightarrow R_k) = occ(CA_i \rightarrow R_k) / occ(CA_i). \quad — (1)$$

$$P(R_k \rightarrow CA_i) = occ(R_k \rightarrow CA_i) / occ(R_k). \quad — (2)$$

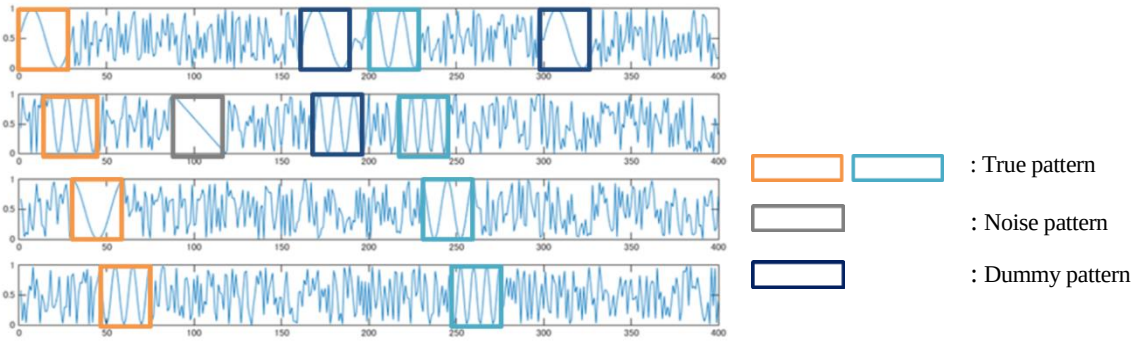

Fig.3: Artificial datasets

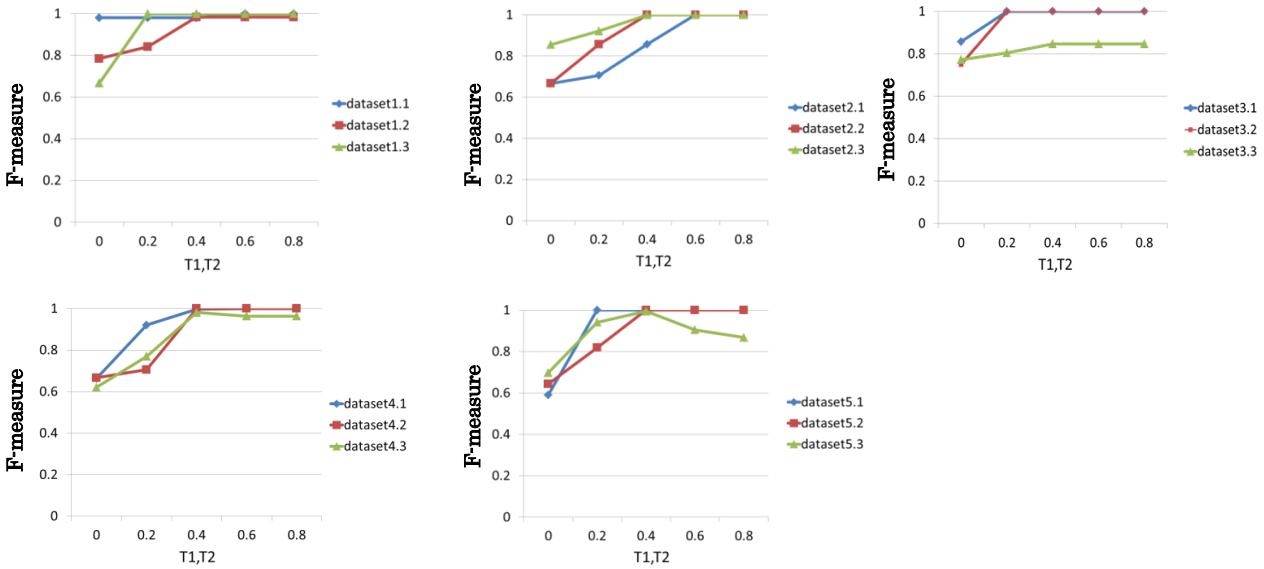

Fig.4: Experimental results

2.3 因果関係のないパタン間のエッジを削除

Fig.2(c)では，因果関係が認められないパタン間のエッジが削除されている．ここでは式(1)の値が閾値 T1 より下回り，かつ，式(2)の値が閾値 T2 より下回るならば，それらの間には因果関係が存在しないと考え，エッジを削除する．なお，本研究では T1=T2 する．

2.4 結果パタンと因果関係のある原因パタンの出力

着目している結果パタンとエッジ接続のある原因パタンを因果関係のある原因パタンとして出力する．

3 評価実験

本研究では，人工的に生成した系列データセット（以下，人工データ）を用いて抽出精度の評価実験を行った．人工データは，4 つの系列で構成され，系列 CA，系列 CB，系列 CC に原因パタン，系列 R に結果パタンが出現するものとする．各系列の長さは 4000 である．

一様乱数によるランダムの系列データに，長さが 30 の原因パタンと結果パタンで構成される 2 種類の連鎖パタン（以下，正解パタン）をそれぞれ 10 個ずつ，長さが 30 の因果関係の無い原因パタン（以下，ノイズパタン）を各系列に 10 個埋め込んだ． 正解パタンの出現間隔を 200 とし，ノイズパタンを埋め込む位置は一様乱数により決定された．

以上の手順により，本実験では以下の 15 種類のデータセットを作成した．dataset1 はパタンの長さを変更した．パタンの長さを dataset1.1 では 10, dataset1.2 では 30, dataset1.3 では 50 とした．dataset2 は正解パタンの出現間隔と各系列の長さを変更した．正解パタンの出現間隔を dataset2.1 では 100，dataset2.2 では 200，dataset2.3 では 300 とし，各系列の長さを dataset2.1 では 2000，dataset2.2 では 4000,dataset2.3 では 6000 とした．dataset3 は正解パタンの出現頻度を変更した．正解パタン出現頻度を dataset3.1 では 10(5×2 種類), dataset3.2 では 20(10×2 種類), dataset3.3 では 30(15×2 種類)とした．dataset4 はノイズパタンの出現頻度を変更した．ノイズパタンの出現頻度を dataset4.1 では 5，dataset4.2 では 10, dataset4.3 では 15 とした．dataset5 はノイズパタンの出現頻度を 20 とし，ノイズパタンの中に正解パタンの一部（以下，ダミーパタン）を挿入し，ダミーパタンの出現頻度を変更した．ダミーパタンの出現頻度を dataset5.1 では 5, dataset5.2 では 10, dataset5.3 では 15 とした．

Fig.3 は，人工データの一部である.パラメータ T1,T2 は 0,0.2,0.4,0.6,0.8 が設定された．正解パタンの抽出精度の評価指標として，下式に示される適合率，再現率，F 値が用いられた：

適合率 = CDP/DDP, 再現率 = CDP / EDP,

F 値 = 2×適合率×再現率 / (適合率 + 再現率)

これらの値は 0 から 1 までの値を取り，大きな値ほど抽出精度が高いことを意味する．

ここで，CDP は正しく抽出された正解パタンのデータ点数 DDP は，本手法により連鎖パタンとして抽出された部分のデータ点数，EDP は正解パタンのデータ点の総数を表している．

4 結果と考察

本稿では，紙面の都合により F 値の結果のみ示す．Fig.4 は，dataset1~dataset5 に本手法を適用したときの F 値のグラフである．

dataset1~dataset3,dataset4.1,4.2,dataset5.1,5.2 では，T1, T2 の値は大きいほど F 値が高くなっており，前手法より大きく上回っている．これは，結果と無関係な原因パタンを連鎖パタンから削除することにより，因果関係のあるパタンで構成される連鎖パタンのみを正しく抽出しているためである．

しかしながら，dataset4.3 と dataset5.3 では，T1,T2 の値が 0.4 のとき，F 値が 1 でピークとなり，それより大きな T1,T2 では，F 値が低下している．dataset4.3 ではノイズパタンの出現頻度が 15 と大きいため，正解パタンと同時刻帯にノイズパタンが出現する確率が大きくなる．これにより，原因パタンと結果パタンの考慮すべきエッジの本数が多くなり，正解パタン内の因果の確率が小さくなるため，本来残されるべきである正解パタンが削除され，F 値が低下したと考える．また，dataset5.3 ではダミーパタンの出現頻度が 15 と大きいため，ダミーパタンの影響により，正解パタン内の因果の確率が小さくなる．よって，正解パタンを正しく検出できなくなってしまうため，F 値が低下したといえる．

また，dataset2.1 と dataset3.3 では，T1,T2 の値が低く設定されたとき，他のデータセットよりも抽出精度が低下している．dataset2.1 のように正解パタンの出現間隔が短い，あるいは dataset3.3 のように正解パタンの出現頻度が多い場合，データセットに出現するパタンが密になるため，ノイズパタンが混入しやすくなることにより，F 値が低下したと考える．

これらより，パタンが密に埋め込まれているデータセットに関しては，T1,T2 の値を大きく設定したほうが高い抽出精度を得られることがわかる．しかし，ダミーパタンが含まれるデータセットに関しては，明確な知見が得られなかったので，今後検討していく必要がある．

5 まとめ

本研究では，異なる系列データに現れるパタン間の因果関係を考慮した新しい連鎖パタンマイニング手法を提案した．多様な人工データに適用させた性能評価実験により，パラメータの設定方法について検討を行なった．結果，パラメータ T1,T2 を大きい値に設定することにより，前手法よりも着目している結果パタンとそれを誘導する原因パタンからなる連鎖パタンを高精度で抽出できることが示された．一方で，ダミーパタンが含まれるデータセットに関しては，明確な知見が得られなかったので，今後検討していく必要がある．

今後は，系列データの性質に応じて閾値を設定する方法の開発を目指す．また，実データに対する適用実験も行い，よりノイズに頑健な手法へと拡張を行っていく．

参考文献

1) 三浦貴大， 岡田吉史 “ノイズを含む系列データの連関パタンマイニングの適用” 2013.
2) Takahiro Miura and Yoshifumi Okada，“Extraction of frquent Association Patterns Co-occurring across Multi sequence Data”， Proc. of IMECS 2012，International Association of Engineers， pp. 452-455， 2012.