



室蘭工業大学

学術資源アーカイブ

Muroran Institute of Technology Academic Resources Archive



因果関係を考慮した連鎖パターンマイニング手法のパラメータの設定

メタデータ	言語: jpn 出版者: 計測自動制御学会 公開日: 2016-02-17 キーワード (Ja): シーケンシャルパターンマイニング, 連鎖パターンマイニング, 因果関係, データマイニング キーワード (En): 作成者: 大久保, 勇輔, 李, セロン, 岡田, 吉史 メールアドレス: 所属:
URL	http://hdl.handle.net/10258/3861

因果関係を考慮した連鎖パターンマイニング手法のパラメータの設定

著者	大久保 勇輔, 李 セロン, 岡田 吉史
雑誌名	計測自動制御学会システム・情報部門学術講演会講演論文集
巻	2015
発行年	2015-11-18
URL	http://hdl.handle.net/10258/3861

因果関係を考慮した連鎖パター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ーン集合である連鎖パター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ー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ー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ーンマイニング, 連鎖パターンマイニング, 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ー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り現れるパターン群（以下、連鎖パターン）を発見する連鎖パターンマイニング手法（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ーン間の因果関係に基づく、新しい連鎖パター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ーン（結果パターン）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ーン（原因パターン）を特定する。パター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（結果パターン）の発生原因（原因系列や原因パターン）を見つけ出すことができるようになることが期待される。

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。

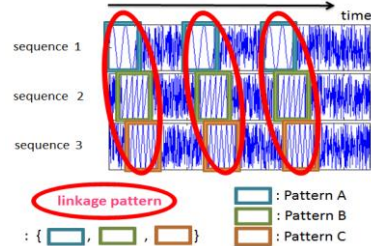


Fig.1: Example of linkage pattern

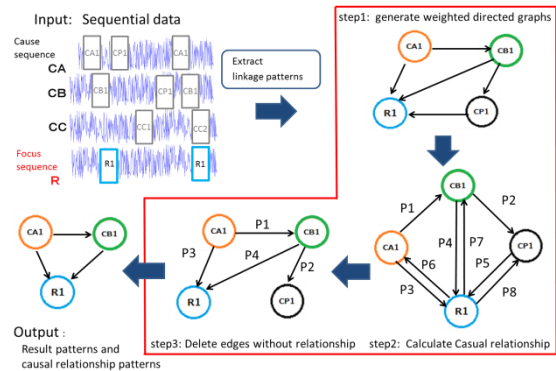


Fig.2: Procedure of this method

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ーンの抽出を行う。本手法は、抽出された連鎖パターンに含まれるパター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に出現するパターンを原因パターンCA_n, CB_n, CC_nとし、着目系列Rに出現するパターンを結果パターンR_xとする。ここで、n, x はパター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ー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ーン間の因果関係の数値をエッジに付与させる。

2.2 パターン間の依存関係の数値化

2.2.1 原因パターンと原因パターンの依存関係の数値化

Fig.2のstep2を用いて原因パター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、各連鎖パター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 原因パターンと結果パターンの因果関係の数値化

次に、Fig.2のstep2を用いて原因パターンと結果パターンの因果関係の数値化を説明する。R_kの出現頻度を $occ(R_k)$ 、各連鎖パター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：

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

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

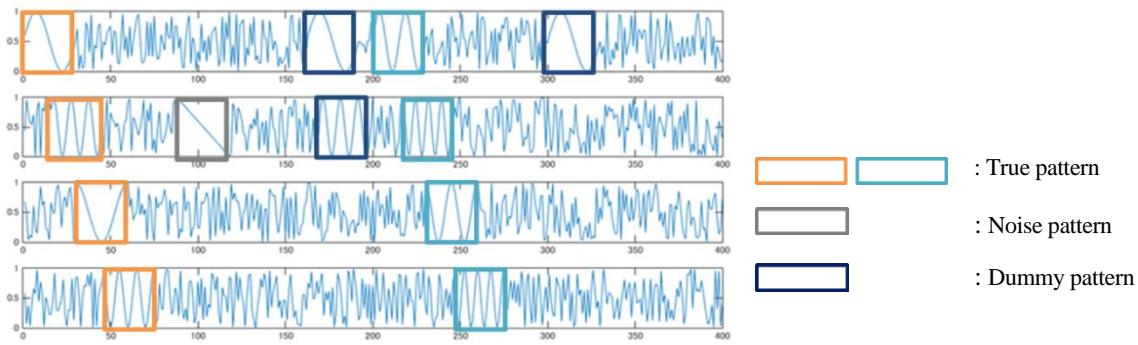


Fig.3: Artificial datasets

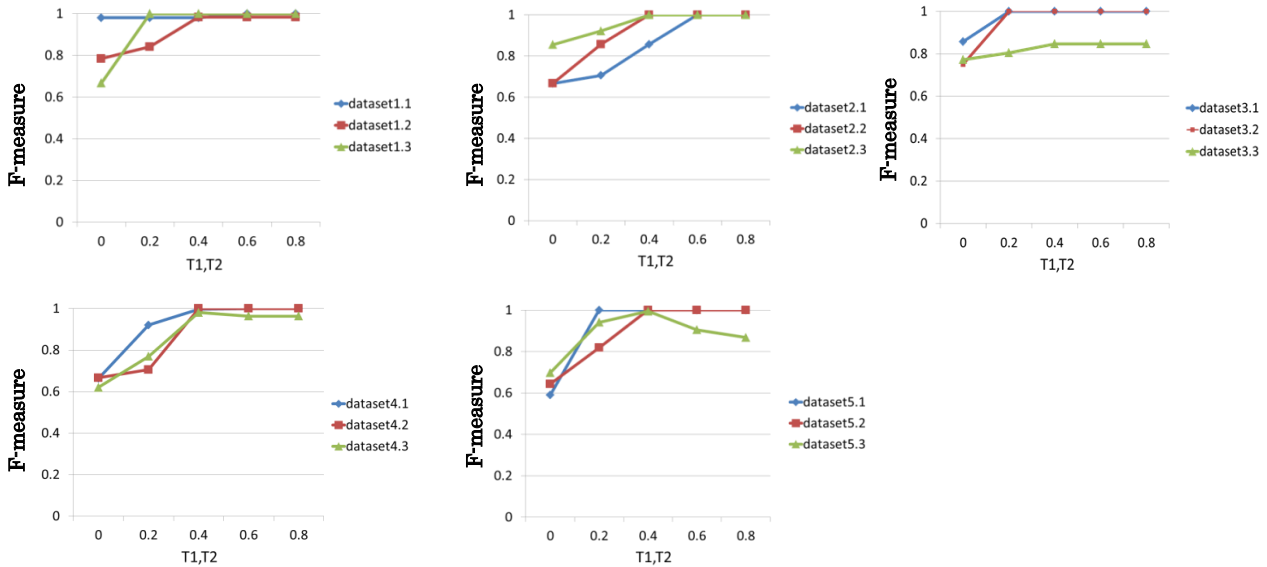


Fig.4: Experimental results

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な系列データに、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, 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：

$$\begin{aligned} \text{適合率} &= \text{CDP} / \text{DDP}, & \text{再現率} &= \text{CDP} / \text{EDP}, \\ \text{F値} &= 2 \times \text{適合率} \times \text{再現率} / (\text{適合率} + \text{再現率}) \end{aligned}$$

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ー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値の結果のみを示す。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ーンを連鎖パターンから削除することにより、因果関係のあるパターンで構成される連鎖パター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、正解パターンと同時刻帯にノイズパターンが出現する確率が大きくなる。これにより、原因パターンと結果パタンの考慮すべきエッジの本数が多くなり、正解パター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ーンが削除され、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、正解パターン内の因果の確率が小さくなる。よって、正解パター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ーンが密になるため、ノイズパター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、パター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、ダミーパター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ため、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ーン間の因果関係を考慮した新しい連鎖パター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ーンとそれを誘導する原因パターンからなる連鎖パターンを高精度で抽出できることが示された。一方で、ダミーパター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ため、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鎖パターンマイニングの適用” 2013.
- 2) Takahiro Miura and Yoshifumi Okada, “Extraction of frequent Association Patterns Co-occurring across Multi sequence Data”, Proc. of IMECS 2012, International Association of Engineers, pp.452-455, 2012.