3A1-1

稀な事象同士の関連性指標
異常値間の関連性抽出のための時系列データマイニング

Discovering relational rules between outliers

金城敬太[*1] 市瀬龍太郎[*2] 相澤彰子[*2] 小暮厚之[*3]
Keita Kinjo , Ryutaro Ichise, Akiko Aizawa, Atsuyuki Kogure
[1]総合研究大学院大学複合科学研究科 [2]国立情報学研究所 [3]慶應義塾大学
[1]Graduate University for Advanced Studies, School of Multidisciplinary Sciences
[2]National Institute of Informatics
[3]Keio University

Abstract In this paper, we propose a method for time series data mining to extract the relevant set of outliers in different time-series. The purpose of this paper is to extract interactions between structural change or anomaly points in each distinctive time series. In our approach, we first identify outliers in time-series. Next, we extract the relationship between the outliers using newly proposed relevant indicators. In our experiments, we applied the proposed approach to simulated and actual data. It is shown that our method can extract useful relational rules.

1. はじめに

今日，ユビキタス技術の普及により，膨大な量のセンサー情報が取得されるようになり、その分析が急務となっている．こうしたなかで，特に時系列データから有効な情報を抽出する時系列データマイニングが注目されており，時系列データのクラスタリングや類似時系列の発見などの問題が扱われている[1]．時系列データマイニングには，異常値の発見というトピックも存在する[2]．これはひとつの系列もしくは複数の系列から異常な値を発見することを目的とした研究である．本研究ではこの中から，ある時系列で異常が起きた結果として，別の時系列にその異常な値の影響が及ぶという問題を扱う．

このような事例は現実でよくみかける．例えば，経営に関する時系列データでは，ある業界のデフォルトなどの異常値が通常は関連のみられない他の業界へ波及することが観測される．このような分布の裾で依存関係が存在している場合，通常の相関係数のように全体のデータで依存関係をみただけでは確認が難しい．また，このような異常なデータの頻度は極めて少なく，一回限りのケースが多く，一回性の出来事同士の関連を抽出するのは統計的な観点からいった場合は困難である．

しかし，リスク管理という観点では，このような異常値の原因特定などは必要である．これらのルールの候補（仮説）をつくることで，仮に一般化はできなくても予め対処することが可能だからである．こうした理由で複数の時系列間での異常な現象の相互関係性の抽出を行うことは意義があろう．

上記で述べたような問題を扱うことは，パラメトリックな統計的方法では扱いにくい問題であるため，データマイニングのようなノンパラメトリックでかつ注目すべき箇所を発見するあらたな方法の開発が必要となる．

以下，2で問題設定し，3で提案手法を述べ，4で検証を行う．

連絡先：金城敬太，総合研究大学院大学複合科学研究科（国立情報学研究所），Tel. 03-4212-2000,kinjo@gmail.com

2. 問題設定

時刻$0, \dots, t, \dots, T$で観察されるひとつの時系列データを

$$N_i = \{n_{i1}, \dots, n_{it}, \dots n_{iT}\}$$

として，k 個の時系列データ

$$NS = \{N_1, \dots N_i, \dots, N_k\}$$

が与えられたときに，そのなかから時系列N_iにおける異常な点データn_{it}と別の時系列$N_{i'}$における異常な点$n_{i't'}$の関係を抽出することが問題である．なお，異常な点というのは，統計的な異常値，すなわち外れ値や，時系列データの構造が変化する変化点などをさす．ここでは，主に前者の場合について扱うが，後者の場合についても変化点の自動検出手法を適用することにより，同様に扱うことができる．

3. 異常値同士の関連抽出法

本研究では「稀な事象でかつ，近接しているもの同士は関連がある」ことを前提とする．稀な事象は，様々な対象で時々おきる．しかし，稀なことがほぼ同時に別の対象でもおきることは極めて珍しいため，そこに関連がある可能性が高いと考える．こうしたルールは人間が因果を捉える際にも使用される．通常は，ポストドックな誤謬・前後即因果の誤謬といわれ，論理的には誤りとされることもある[3]．しかし，1で述べたように一回性の出来事を扱う場合や，リスク管理といった応用の場面においては利用可能な考え方である．その妥当性や具体的な考え方については後で議論する．

上記の「稀な事象でかつ，近接しているもの同士は関連がある」という前提に基づき，提案アルゴリズムは以下で構成される．

Ⅰ．時系列データの予測
Ⅱ．一系列内の異常値の計算
Ⅲ．近接性を用いた異常値同士の関連性の計算

まずⅠ．では，観測時にノイズを含む時系列データから時系列の値の推定を行う．推定にはカーネルをもちいた近似方法を用いる．別の自己回帰モデルなどの時系列モデルも利用可能であるが，ここでは簡便に考えるためにこの方法を用いた．

次にⅡ．では，一系列内の各データの異常度の計算する．推定された真の値とデータがどれくらい離れているかを計算し，その確率値をもとめる．

最後にⅢ．は，近接か否かを考慮にいれた異常値同士の関連性を提案した指標を計算し，最終的に関連が高いものを順に並べることで異常値同士の関連の抽出を行う．こうすることで，変動する時系列内から発見しづらい，異常値同士の関係を自動的に抽出し，その評価を可能としている．また，これらは時系列データが大量にあるときにも使用が可能である．ただし，組み合わせが膨大になるため，計算範囲をしぼる必要上から上位のいくつを異常値とするかの設定が必要となる．

本研究の最も中心的な提案はⅢである．

3.1 時系列データの予測

はじめに時系列データの予測を行う．

ここではノンパラメトリックな回帰を単純に利用する[4][5]．まず，入力データを下記のモデルで表現できると考える．

$$f(t)=\sum_{i=1}^{T} W_i(t)\, n_i$$

$$W_i(t)=K(t-i/h)/\sum_{k=1}^{T} K(t-k/h)$$

f を時間 t に対する関数として，T はサンプル数，Kはカーネル関数を表し，hはバンド幅を表す．これは一般に Nadayara-Watson 統計量とよばれる．入力データ NS のそれぞれのN_iデータの値に対して予測値を求める．カーネル関数としては，ここではガウスカーネルを用いる．

またバンド幅の選択，すなわちカーネルの幅を規定するhについて次に考えたい．これら，カーネル幅の設定は任意ではなく，より真の値を推定していると考えられるものを選択する．このバンド幅はのちに用いる関連性指標にも関係する．実際に最適なバンド幅を選択するためにいくつか方法が提案されており，GCV（General cross-validation）などが存在する．なお，これらとは別にノンパラメトリックな方法を用いて時間的な依存も扱う時系列分析の方法も存在している[6]．

3.2 一系列内でのデータの異常度の計算

次に上記のような値をもとに，一時系列N_i内で個々のデータが予測値n'_{it}からどれくらい離れているかをもとにして異常値を計算する．

具体的には誤差が，正規分布に従うとし，これを利用する．各データの出現する確率をデータとの誤差の分散をもとに計算した正規分布を利用し，さらにその確率値の逆数を異常度 I と定義する．すなわち，以下のように定義される．

$$I_{it} = 1/P(|n_{it}-n'_{it}|)$$

n'_{it}は推定値で，ただし，P は誤差の分散を用いた正規分布 X の P(x)=Prob(x≦X)とする．

3.3 近接性を用いた異常値同士の関連性の計算

つづいて複数の系列間での，異常点同士の関連をその異常度およびデータ間の距離をもとにした関連性指標を計算する．

データ系列N_1とデータ系列N_2が与えられた場合のその内部の異常点n_{1t} $(1 \leqq t \leqq T)$およびn_{2t} $(1 \leqq t \leqq T)$の関連性指標の一般系を以下のように定義する．

関連性＝ (データn_{1t}の異常度)
×(データn_{2t}の異常度)
× β (n_{1t}とn_{2t}の近接性)

すなわち関連性の値は，稀なものどうしが近接しておきた場合に大きくなる．別の見方をすると，遠くの点も含めた範囲内で考えれば，稀な事象が共起する確率は極端に低くはならないため，共起したからといって必ずしも関連があるとはみなせないことになる．なおこれら異常度と近接性の間は重み β で調節を加えることもできる．

上記で近さを示す近接性という概念を導入したが，近接性については，大きく以下の3つの種類があると考えられる．

①時間的近接性

時系列での近接性をはかる場合，時間的に近いものには重みを与えて計算を行えばよい．例えば，系列N_1とN_2の異常値がそれぞれn_{1t}とn_{2t}であったとする．それぞれの時間が $t=t_1$,$t=t_2$ であった場合，t_1とt_2の時間的に近い度合いが時間的近接性となる．

ただし，それぞれの系列N_1とN_2において，それぞれが持っている関係構造をみていく必要がある．例えば，N_1は直近のデータとの依存が強いのに対し，N_2は距離が離れたデータの影響も受け易いなども考慮する必要がある.本研究ではこうした各系列の特有の構造を時間的近接性に組み込むために，3.1 で用いた推定で用いたような重みを用いている．すなわち，$t_1<t_2$としたときn_{1t}のn_{2t}への影響をかんがえるとn_{2t}からみたt_1時点の重みをその時間的近接性の値として用いる．

なお，これ以外にも近接性の導入方法は考えられ，例えば単純に時間的な距離（$|t_1-t_2|$）を直接用いるなど様々なものを想定できる．

② 空間的近接性

これは，「稀な現象が発生している距離が近いので意味がある」ということを示すものである．でありユークリッド距離など利用が出来る．ネットワークなどの構造をもったデータの場合は，経路長を用いることも可能である．例えばN_1という系列が持つ空間的な属性と，N_2の持つ空間的属性との距離を計算する場合を例にあげると，N_1とN_2に付随する空間的座標を仮に(d_1, e_1),(d_2, e_2)とするとこれらのユークリッド距離などを用いることができる．

③カテゴリ的近接性

時系列データのクラスがあるのであれば，その概念の近さを用いることも出来ると考えられえる，これをカテゴリ的近接性と定義する．つまり，異常なものでかつ意味が近いものであれば関連するという考えである．例えば，家計調査などで支出の時系列データを扱う場合には，その支出項目間の距離を概念階層の距離（たとえば、項目の階層を木構造とらえてそのパスの長さ）をもとに計算することも可能であるし，その付随する属性間の距離をユークリッド距離などを用いて計算することも可能である．形式的に表現すると，空間的近接性とほぼ同様に定義され，属

性を仮に(d_1, e_1),(d_2, e_2)とするとこれらの距離で計算できる. ただし, この近接性については必ず成り立っているわけではない. たとえば距離の離れたカテゴリ同士の間に因果関係が成立することもあるため, 探索空間が広くあらかじめ特定のカテゴリ内での関連を調べるのに限定したいときに利用できると考えられる.

以上の3つの指標は, 組み合わせて使用することも可能であるが, このうちもっとも一般的なものが時間的近接性であろう. そこで本研究では, 時系列に焦点をあてて抽出法を説明する. ただし, 他の近接性についても, 時間的近接性に基づき計算した値に追加して考えることでより限定した異常値間の関連性を導くことができる.

3.4 抽出アルゴリズム

これまで紹介した指標に基づき関連のある異常値のペアを抽出する方法を下記に簡単に述べる.

まず, 各系列で推定を行い, 次にその一系列内の異常値の計算, そして異常値同士の関連性指標の計算を行う. 異常データについては, 上位 p 個に絞ることで候補を削減することができる. ただし全探索ではない. 出力は, ある系列上の異常データとある系列上の異常データとその関連性の指標の組になる.

D=outliermining(NS,p)
1. 時系列集合 NS よりN_iを抽出
2. N_iの予測を計算N'_i
3. 異常度データI_iをN_iとN'_iにより計算
4. 1,2,3 をすべての i に対して実行
5. 全異常データ I の異常度をソートし, 上位 p 番目の値を計算
6. p 番目の値より大きい異常データセットを O に保存
7. O より関連性指標を計算, 算異常データ対セット D に保存
8. 異常データ対セット D を関連指標の高い順にソート
9.D を出力

図1.抽出方法

4. 実験

この章では, 提案した手法を用いて実験を行う.
テストデータに対して実験を行ったあと, 実データに対して適用を行い, その結果の解釈をする.

4.1 テストデータ

以下の手順で, 異常値がある複数系列のデータを用意する. まず下記で示す2系列の VAR(ベクトル自己回帰モデル)に従う系列を用意し, そのなかに異常な値をいれる. その波及を実際に, 本手法で抽出できるかどうかの実験を行う.

VAR とは下記のような式であらわされるモデルである.
X をベクトルとすると下記で表現され,t は転置を示す.

$$X_{t+1} = \alpha X_t + U_t$$
$$E(U_t)=0,\ VAR(U)=E(U,t(U))=\Sigma$$
$$Cov(U,t(U))=E(U,t(U))=0$$

このモデルに従う系列をシミュレーションを行い, データを作成し, その途中一系列 x の時点 t に異常なインパクト I を与え, 結果として別の系列 x'に波及するということを想定する.

このとき, x における異常なインパクトと, 別の系列 x'におけるインパクトを本手法によって抽出できるか否かをテストする.

評価基準としては, 検出したルールと実際に事前に設定した異常値であったルールがどれくらい検出できているのかの比率を計算する. それにより, 実際にどの程度の値であれば本手法においで検知できるか否かを調べることができる.

全データは 200, パラメタ α すなわち X 自身への係数は 0.3, 別の系列からの係数は 0.5 とし, また初期値はそれぞれ 0.5 に設定した. この系列に t=100 の時点で追加するインパクト I を変動させたときに, ルールの上位 10 のうち, 検出できたか否か(t=100±5 でルールを検知しているかどうか)をテストする. これらを 50 回繰り返し, このうち検出できた比率を導出する. アルゴリズムのパラメタは t=20 とし, n は 10 とした. I は1から10までの大きさで評価を行なった. また, 今回, 関連性指標の重み β = 1を利用した.

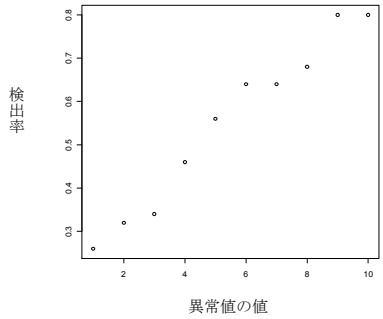

図2 異常値の値(横軸)と検出率(縦軸)

結果は, 図2のようになった. 異常値が1のときは 0.2 の検出であったが, 異常値が高くなるにつれて, その抽出率が高くなった. 200 のデータのとき 19900 個の組み合わせが想定される. このうち, インパクトのあった t=100±5 でみた場合は, 55 が検出に必要なルールとなる. こうした極めて稀なルールが検出できていることになる.

4.2 実データでの実験

次に実データでの実験を行った.
本研究では, 家計調査による米の消費量の時系列推移と米の消費者物価指数の時系列推移の関連を分析し, その後, 得られた結果についての解釈を行う.

1990 年から 2009 年までの一つの世帯あたりの米の購入数量と, 米の消費者物価指数の推移をみてみよう.

米の購入数量は, 年々減少していることがわかる. それと同時に価格も全体として低下してきていることがみてとれる. これらは需要と供給との関連からも説明がつくと考えられる. こうしたなか, 1994 年と 2004 年については, 物価の上昇が一部みうけられており, それと同時に購入数量の低下も観測される. (図3).

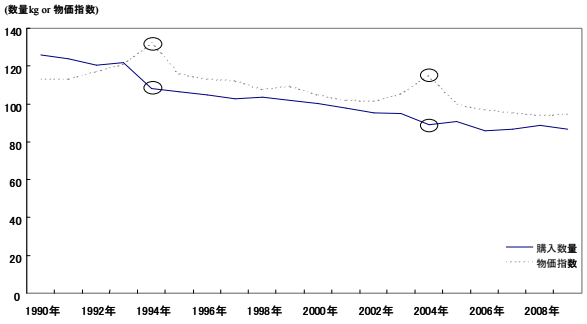

図3.二人以上世帯(農林漁家世帯を除く)の米の消費数量と物価指数

続いて，このデータに対して今回提案した手法を用いて得られた結果を示す．今回，関連性指標の重み β ＝1として計算した．提案した指標の値が高い上位4つは以下のものである(表1)

	事象A		事象B	
値	年代	系列	年代	系列
8.397	1994年	数量	1994年	価格
7.668	2006年	数量	2006年	価格
7.674	1993年	数量	1993年	価格
7.687	2004年	数量	2004年	価格

表1　抽出ルール

以上のように，1994 年の価格と数量に関するルール，そして 2004 年の価格と数量に関するルールが得られた．この場合，価格の上昇により購入数量がやや減少していると考えられる．つまり，全体としては数量の低下に伴った価格の低下として関連があるものの，一部だけ価格があがったために減少するといった別の関係が存在していることがわかる．このようにある極端な値をとった場合に別のメカニズムが働いている場合にルールを抽出するのは難しい．ただし，2006 年や 1993 年など別のルールも得られており，これについては特徴のデータが大きく動いた前後の特徴を拾ったために抽出されたものと考えられる．こうしたルールの解釈や対応については今後の課題としたい．

5. 関連研究

稀ではあるが，意味のあるデータというものを扱った研究としては，相関ルールにおける lift 値などがある[8]．これらは仮にまれであっても共起するのであれば，関連が高いとみなすものであるので，本研究の意図とも近い．ただし，共起に加えて近接性という観点を導入したことが本研究でのひとつのポイントともいえる．また文書検索の分野でよく用いられる指標として TF/IDF などが知られている．これは，ある文書だけで頻度の高いものを重要であるとみなすというものであるので，今回のケースとは設定が異なる．

それ以外で扱ったような分布の裾における現象を取り扱う事例としては，コピュラで表現してシミュレーションなどを行っている研究がある[9]．

しかし，稀な事象間の関係性を扱った研究はなく，さらに統計的な意味では除外されがちであるルールについての取り扱いを行ったということで本研究は意義がある．

また，こうした計算のために近接性という概念を導入することでより，精度を挙げることが可能になった点でも意義があると考えられる．

6. まとめ

今回，近接性ということに注目し，関連指標を開発し異常な値同士の関連性を抽出するということに対して有効性を示すことができた．

ただし，この手法については，近接性としてどの値を用いるか，その重み β はどうするか，距離をどのように定義するかなどの選択の問題も存在している．一般には，関連する異常値同士が存在するデータを与えて，その検証能力をもって，どの推定値を用いて，さらにどの規準で異常とするか，そして，時間的距離などの近接性を定義していくことが可能であろう．

問題としては，異常値が比較的近い場所で大量に発生している場合，その付近のルールを抽出してしまうことがある．今回の研究では，抽出するルールの個数を限定したり，同じ系列の組でのルールは限定するなどを行い対処したが，今後はこうした不要なルールをいかに排除するかなどの方法論を考えていく必要があると考えられる．

今後の可能性としては，これら特殊な状況での時系列データの相関を移動相関としてみていくことも考えられる．こうした方法との比較や違いも検討していく必要があろう．

本研究では，工学的な必要性から問題を設定していったが，こうした事態は工学的な必要性のみならず，人間における推論でも重要な役割をはたしていると思われる．人間の場合も，一度しかデータが得られないケースがあり，それに基づいてなんらかの因果推論を行うということがある．これらはアドホック論理として扱われているものである．こうした方法は多くの場合，過度の一般化であったり，誤謬の一種であるとみなされるが，なんらかの仮説を生成したり，情報が非常に限られている中で推論を行わなくてはいけないという状況下においては重要な方法とも考えられる．特に大きなリスクを伴った事態においては，こうしたルールは仮に稀であっても重要であると考えられる．

謝辞

本研究の研究は，国立情報学研究所佐藤健教授のアドバイスをもとに実現しましたことをここに感謝します．

参考文献

[1] Mark, L., Kandel, A. and Bunke, H.: *Data mining in time series databases*, World Scientific , 2004.

[2]藤木稔明 南野朋之 鈴木泰裕 奥村学,document stream における burst の発見 情報処理学会研究報告．自然言語処理研究会報告, Vol.2004, No.23(20040304),pp. 85-92, 2004.

[3]佐々木憲介, マルサスにおける帰納と演繹, 經濟學研究 ,The economic studies , 45(4), pp.35-48, 1996.

[4]竹澤邦夫 みんなのためのノンパラメトリック回帰 吉岡書店, 2007.

[5] 小暮厚之 ノンパラメトリック平滑化法による信用リスクの測定, S-PLUS ユーザーカンファレンス, 2002.

[6] Siegfried Heiler, A Survey on Nonparametric Time Series Analysis, Finance 9904005, EconWPA, 1999

[7] Siegfried Heiler, Nonparametric Regression, Locally Weighted Regression, Autoregression, and Quantile Regression., Chapter 12 in: A Course in Time Series Analysis., Daniel Pena, George C. Tiao, and Ruey S. Tsay. John Wiley & Sons Inc, 2001

[8]岡田孝 元田浩, 相関ルールとその周辺, オペレーションズ・リサーチ : 経営の科学 47(9), 565-571, 2002.

[9] 戸坂凡展 吉羽 要直, コピュラの金融実務での具体的な活用方法の解説, 金融研究第 24 巻別冊第 2 号, 2005.