D-020

特徴検知に基づくストリーム中の因果関係の監視

Monitoring of Causal Relationships on Data Stream Using Characteristics Detection

山原 裕之† 島川 博光†
Hiroyuki Yamahara Hiromitsu Shimakawa

1. はじめに

株や為替の経済指標，火力発電所のデータのように，多くの分野においてセンサなどから膨大な時系列ストリームデータが収集されている．一般に，専門家は時間経過に伴う状態遷移の中に，特定の特徴を持つ状態遷移パターンに適合する状態遷移が現れるかどうかを監視する．

時系列データの中から特徴を持った状態遷移パターンを見つけ出す手法は，これまでにも研究されている．Keogh らの状態遷移間の距離を計測してその類似性を判定する手法[1][2]は，期間ごとに状態遷移の特徴を直線で近似して表現しており，これは類似性の判定の高速化を目的としている．ゆえに距離の計測による手法は，状態遷移が持つ期間ごとに注目すべき値が異なる特性を柔軟に表現できない．これに対して，本システムにおける特徴パターンを用いた検知ではこれを柔軟に表現できる．

本論文では，これに代わる手法として期間分割を用いた手法を提案する．本手法は，状態遷移を複数の期間の連なりであると考え，期間ごとに特性を指定することで状態遷移の特徴を柔軟に表現する．本手法では，専門家が過去の状態遷移のサンプルを複数個用いて，監視対象のデータの中から見つけたい状態遷移のパターンを感覚的に指定する．状態遷移を複数の期間に分割する際の各期間の長さはサンプルごとに異なっているため，状態遷移パターンにおける各期間の長さは一定の範囲を持つ．また，この手法を ECA ルール[3]として実装した能動型ストリームデータベースシステムを提案する．本システムは，関連する状態遷移パターン間の関係をルールとして表現し，複数のルールを並行して実行することで，ルールが成立すれば即座に，能動的にユーザに働きかける．

ストリームデータを監視する場合，現在の値だけではなく，現在までの値の状態遷移が重要な意味を持つ．すなわち，現在の値がどれだけ続いているか，あるいはある条件がどのタイミングで満たされなければならないかといった時間の概念が必要となる．しかし，Lee らの研究[4]に見られるように，従来のルールベースシステムの研究ではルールの実行に関して時間の概念は扱っていない．

Marcuello らは，データ予測を用いた処理の高速化について考察している[5]が，この手法ではストリームデータ内の特徴ある状態遷移を見つける上での問題を解決することは困難である．隣接する２つの期間の特性を両方満たしているデータが現れた場合，そのデータがどちらの期間に含まれるかは，それ以降のデータを見なければ判断できない．ストリームデータでは，このような複数の可能性を常に並行して検討しなければならない．検討すべき可能性が増えると，システムの負荷が高くなる．

本論文で提案するシステムは，ストリームデータの認識に特有の時間の概念を扱っており，負荷上昇の問題を解決する機構を持つ．

† 立命館大学理工学研究科

2. 特徴パターンと特徴パターン間の因果関係

2.1 特徴パターンと期間特性

時系列データを監視する場合には，監視対象のデータ項目値がどのような状態遷移をしているか，そして特定の特徴を持つ状態遷移パターンが現れているかが見極められている．本論文における，時系列データ中に現れる特定の状態遷移パターンの特徴の捉え方を，図１で説明する．図１の横軸は時間を示し，縦軸は変量 v を示す．図１の状態遷移は，感覚的には３つの期間に分けて捉えることができる．

1. 期間 S_1 ： $v_1 \leqq v \leqq v_2$ ： 横這い
2. 期間 S_2 ： 傾き $\geqq \delta$ ： 上昇
3. 期間 S_3 ： $v_3 \leqq v \leqq v_4$ ： 横這い

３つの期間は，それぞれの期間における状態を表現するために注目すべき値が異なる．S_1 では値が v_1 から v_2 までの範囲に収まっているかどうかに注目すべきである．S_2 では値の上昇率に注目すべきである．S_3 では値が v_3 から v_4 までの範囲に収まっているかどうかに注目すべきである．S_1 と S_3 は同じ横這いであるが，値が収まるべき範囲に違いがある．このように本論文では，特定の特徴を持つ状態遷移パターンを，注目すべき値が異なる複数の期間の連なりと捉え，これを"特徴パターン"と呼ぶ．特徴パターンは，時系列データ中に周期的に現れるのではなく，一過性の状態遷移の特徴である．ある状態遷移に関して，各期間での注目すべき値が一定の条件を満たしていれば，その状態遷移は特徴パターンに適合している．この期間ごとの条件を"期間特性"と呼ぶ．各期間の期間特性は項目値条件と時間条件の２種類で構成される．項目値条件は，監視対象のさまざまなデータ項目値に関する制約条件であり，データ項目値に関する不等式で表現する．時間条件は，各期間の期

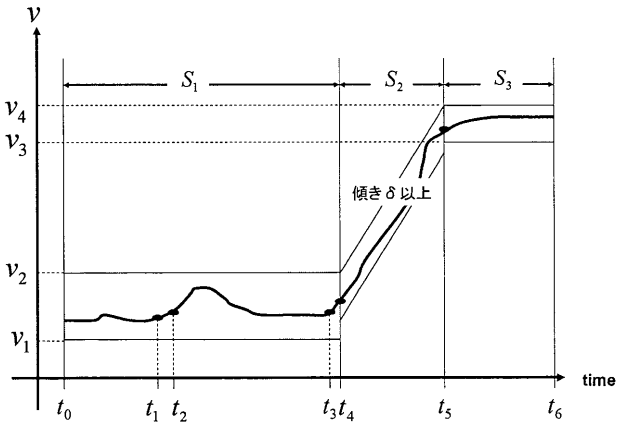

図１： 特徴パターン

間長に関する制約条件であり，最短期間長と最長期間長の２つで表現する．ある期間の項目値条件を満たしている時間の長さが最短期間長以上かつ最長期間長以下であれば，条件は満たされる．

特徴パターンとその各期間および期間特性は，対象とする分野の専門家の知識や経験から獲得できるものとする．また，時系列データの中から特徴パターンに適合する状態遷移を見つけ出すことを"検知"と呼ぶ．

2.2 因果関係を表現するルール

監視対象のある時点での状態遷移は，過去の何らかの状態遷移の影響を受けていると考えられえる．たとえば為替変動では，ある時点での状態遷移から判断されたさまざまな売り買いの結果，その後の状態遷移が起こったと考えられる．そこで本論文では，関連する複数の特徴パターン間の関係を因果関係と捉え，因果関係をルールとして表現する．

図２にルールの構成例を示す．ルールには，検知すべき状態遷移の特徴パターンとそれらを検知すべき順番を指定する．先頭に指定された特徴パターンに適合する状態遷移を"原因"と呼ぶ．それに対して，ルールが成立したときの行動を"結果"と呼ぶ．原因が現れてから結果が引き起こされるまでの間に指定された特徴パターンに適合する状態遷移は，ルールが成立する前触れだと捉えることができる．これを"前兆"と呼ぶ．ルールには，原因が現れてから前兆が現れるまで，また前兆が現れてから次の前兆が現れるまでの間に存在するタイムラグを指定する．各特徴パターンに適合する状態遷移は，指定されたタイムラグが経過したと同時に現れるとは考えにくく，実際には多少の時間のずれが存在すると考えられる．そこで検知のために許される猶予時間の長さを指定する．これを"検査期間"と呼ぶ．指定した検査期間の間は何回でも繰り返し検知可能であるとする．原因と前兆を全てルールどおりに検知すればルール成立とみなす．

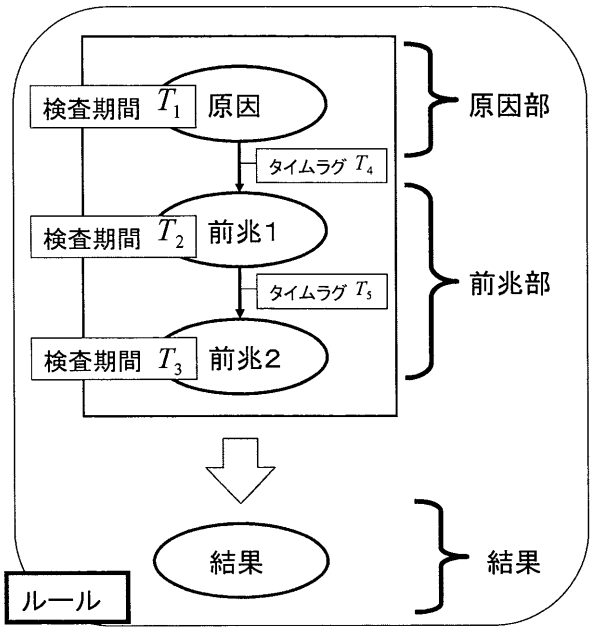

図２： ルール構成例

3. 能動型ストリームデータベースシステム

3.1 検知とルール実行

本論文では，センサなどから非同期で連続的に送られてくる膨大な量の時系列ストリームデータの監視を行い，特徴パターンに適合する状態遷移の検知を繰り返すことでルールが成立しているか否かを判定し，成立している場合には即座に能動的にユーザに働きかける能動型ストリームデータベースシステムを提案する．

本システムでは，ある特徴パターンに適合する状態遷移を検知するために，最新の値を取得する度に項目値条件を判定し，項目値条件が満たされ続けている時間の長さを時間条件で判定する．期間特性を満たしているかどうかは最新のデータを監視することで判断できるので，過去のデータを扱うためにデータベースへアクセスすることによる負荷の増大を回避できる．時間条件が満たされれば検査中の期間の期間特性が全て満たされるので，次の期間の判定を進める．あるいはその期間が特徴パターン中で指定された最後の期間であるなら，検知したとみなす．また，ある原因または前兆が現れてから次の前兆が現れるまでの間のタイムラグ，それぞれの検査期間をルール実行時に判定する．

3.2 能動性に起因する過負荷

ストリームデータを監視するシステムは，監視対象の状態に応じて即座にユーザに対して自動的に働きかけるDBMS-Active, Human-Passive[6]の性質が要求される．しかし，そのためにシステムには大きな負荷がかかってしまう．

各原因と前兆は検査期間を指定されている．ルール実行時には検査期間の間であれば何回でも検知が可能であるから，検知回数に応じて実行中のルールが成立する可能性の数が増えていく．システムは複数の異なるルールを並行して実行すると同時に，このようなルール実行中に増えた複数の可能性を個別に並行して実行しなければならない．同時に実行するルールの数が増えれば，必然的にシステムの負荷が大きくなってしまう．

また，各期間の期間長が一定の幅を持っており，それらが重なることによって，ルールが多重起動する可能性がある．図１のグラフをある監視対象の状態遷移とする．横軸は時間を示し，縦軸は監視対象の変量 v を示す．この状態遷移は３つの期間 S_1, S_2, S_3 とそれぞれの期間特性で構成される特徴パターンとして検知されると仮定する．システムが時刻 t_0 から 時刻 t_1 までに取得したデータは S_1 の項目値条件を満たしており，時刻 t_1 で S_1 の時間条件を満たしたとする．また，時刻 t_2 に取得した次の最新データが S_1 と S_2 の両方の項目値条件を満たしていたとする．t_2 の一定時間後までデータを取得すれば，t_2 から始まる上昇は一時的なものでありその後下降するため，時刻 t_2 からの期間を S_2 と認識することは不適切だと判断できる．しかし，t_2 の時点ではそれを断定できないため，t_2 からも S_1 に属するという可能性と t_2 からは S_2 に属するという可能性の両方を並行して検討していく必要がある．どちらか一方の可能性を先に検討し，その結果がわかってからもう一方の可能性を検討していたのでは，監視対象の状態に即座に対応することができない．どちらの可能性も正しく，結果としてルールが多重起動するおそれもある．同じことが時刻 t_4 でもいえる．このような場面がいくつも存在すれば，常に複数の可能性を並行して検討しなくてはならず，必然的に

システムの負荷が大きくなってしまう．

3.3 同一視期間

本システムは，能動性に起因する過負荷を抑える機構を持っている．

各ルールの特徴パターンごとに期間 d を考える．ある特徴パターンに適合する状態遷移を検知した後，期間 d の間は再びその特徴パターンに適合する状態遷移を検知したとしても，それをルールの実行に反映させず検知とみなさない．このような d を同一視期間と呼ぶ．その働きは２種類に分けて考えることができる．

第１に，異なるタイミングで検知した本来同じ特徴パターンに適合する複数の状態遷移を，同一とみなす働きがある．仮に，同じ特徴パターンに適合する状態遷移を５回，短期間に検知したとする．この特徴パターンがルール中で原因に指定されていれば，短時間に５つのルール実行が開始されることになり，負荷が大きくなる．同一視期間 d により，５回の検知のうち実際にルールの実行に反映されるのは最初の１回のみとなり，負荷の増大を回避できる．ただし，本来異なるものを同一視することでルール実行上の精度は多少低下する．

第２に，一定の幅を持つ各期間の期間長が重なることによるルールの多重起動を防ぐ働きがある．図３をある監視対象の状態遷移を示すグラフとする．横軸は時間を示し，縦軸は変量 v を示す．また，時刻 t_0 から時刻 t_6 までの間に，期間 S_1，S_2 で構成される特徴パターンに適合する状態遷移を次のように３回検知したと仮定する．

検知１：S_1 [時刻 t_0 から時刻 t_1]，S_2 [時刻 t_1 から時刻 t_4]
検知２：S_1 [時刻 t_0 から時刻 t_2]，S_2 [時刻 t_2 から時刻 t_5]
検知３：S_1 [時刻 t_0 から時刻 t_3]，S_2 [時刻 t_3 から時刻 t_6]

これは開始時点が同じである本来同一の状態遷移に対して，複数の可能性を検討しているといえる．ゆえにこの場合，３回検知したと捉えることは誤りであり，全て同一の，すなわち１回の検知とみなさなければならない．可能性の数は削減できないが，複数の可能性が同じ特徴パターンに適合する場合にこれらの可能性を全て同一とみなすことはできる．同一視期間 d により，検知２と検知３は検知１と同一視されて検知１だけがルールの実行に反映されるため，本来同一のものを間違って異なるものと認識することによる負荷の増大を回避できる．

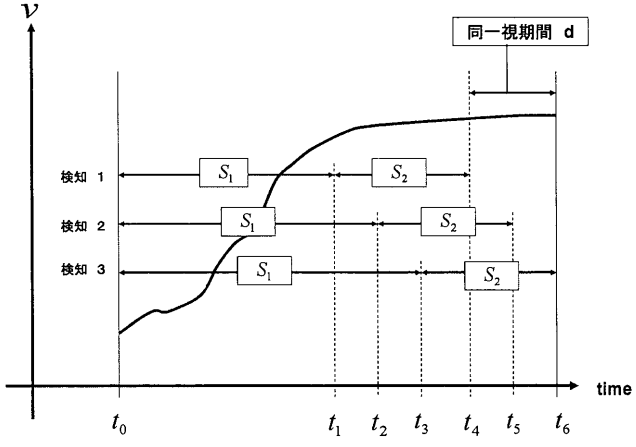

図３： 同一視期間 d

3.4 ルールの実行系

図４にルールの実行系の概要を示す．

トリガスレッドはルールを実行し始めるきっかけを知り，ルールスレッドを生成する役割を持つ．トリガスレッドは全てのルールの原因を検知しようとする．そのためにマネージャスレッドに検知依頼を行う．原因が検知されればマネージャスレッドから検知通知を受け，それをきっかけに検知した原因に対応するルールを実行するためのルールスレッドを生成する．この仕組みは，ECA ルール[3]の E-C Coupling における separate モードと同じである．

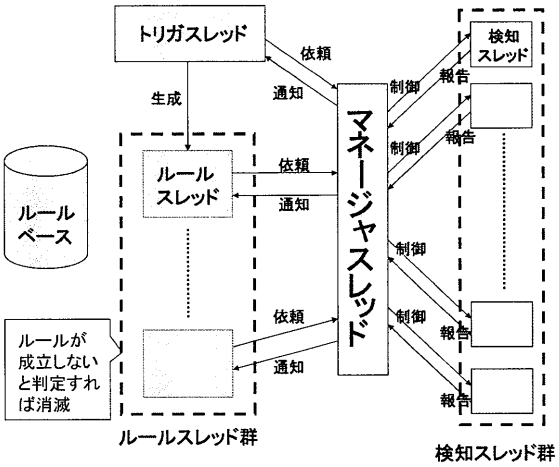

図４： 実行系の概要

ルールスレッドはルールを実行する役割を持つ．トリガスレッドと同様に，前兆を検知するためにマネージャスレッドに検知依頼を行い，検知通知を受ければ次の前兆の検知を依頼する．次の前兆がなければルール成立とみなす．ルールが不成立と判断すれば自スレッドを消滅させる．

検知スレッドは，時系列ストリームデータを監視しながら特徴パターンに適合する状態遷移を見つけ出す役割を持つ．１つの検知スレッドは１つの特徴パターンについて専門的に検知する能力を持つ．検知スレッドは，マネージャスレッドからの制御を受けて動作を開始し，要求された時間の間だけ動作する．担当する特徴パターンに適合する状態遷移を検知すれば，マネージャスレッドに検知報告を行う．

検知管理スレッドはトリガスレッドおよびルールスレッドと検知スレッドの間に位置し，これら２つの働きを結びつける役割を持つ．依頼者であるトリガスレッドおよびルールスレッドから検知依頼を受け，その内容に応じて検知スレッドを動作させる．検知スレッドから検知報告を受けると，その検知を依頼していた依頼者に検知を通知する．また，同一視期間を有効にするために，依頼を通知グループと非通知グループに分別して管理する．

本システムの実行系は，ルール実行の役割と検知する役割を分離している．２つの役割を１つのスレッドに持たせると，複数のルールスレッドで同じ特徴パターンに適合する状態遷移を検知しようとする場合，個々に検知を行うことになる．本システムは，２つの役割を別々のスレッドに持たせ，このような冗長な動作を行わない．また，ある検知スレッドによる検知はその検知の依頼者のみに通知することで動作するスレッド数や条件の判定回数を最適化し，

実行時間を削減している．Lee らの ARS，IRS によるルールの最適化[4]も同じ考え方である．

火力発電所の起動時の TB 弁の開度を監視データとして，ある年度の全 45 個のサンプルデータからランダムに抜き出したサンプル 10 個を用いて特徴パターンを作成した．それを用いて，本システムを別の年度の監視データに適用し，期間特性を用いた検知手法の評価を行った．その結果，検知すべき 46 箇所の状態遷移のうち 40 箇所を正しく検知し，認識率 86.96%を示した．

4. おわりに

本論文では，監視対象の膨大な量のデータの状態遷移の中から目的の状態遷移パターンを見つけるための手法として期間分割を用いた手法を提案した．本手法では，状態遷移パターンを期間の連続と捉え，期間ごとの期間特性を持つ特徴パターンとして定義することで，その特徴を柔軟に表現することが可能になった．関連する複数の特徴パターン間の関係を因果関係と捉え，ルールとして表現した．また，能動型ストリームデータベースシステムを提案した．本システムでは，複数の可能性を並行して検討することによって，必然的にシステムの負荷が高くなるが，同一視期間を用いて負荷の上昇を抑えた．

火力発電所のデータに対してシステムを適用して評価を行ったところ，検知すべき状態遷移のうち 86.96%を検知した．今後，より多くのさまざまなデータに対して実験を重ねて，システムの評価を行う．

参考文献

[1] E. Keogh and P. Smyth. A Probabilistic Approach to Fast Pattern Matching in Time Series Databases. Proceedings of KDD 1997: pp.24-30.

[2] E. Keogh, K. Chakrabarti, S. Mehrotra, and M. Pazzani. Locally Adaptive Dimensionality Reduction for Indexing Large Time Series Databases. Proceedings of ACM SIGMOD, pp.151-162, 2001.

[3] D. McCarthy and U. Dayal. The Architecture of An Active Database Management System. In ACM SIGMOD International Conf. on Management of Data, pp.215-224, 1989.

[4] Y.-H. Lee and A.M.K. Cheng. Optimizing Real-Time Equational Rule-Based Systems. IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL.30, NO.2, pp.112-125, FEBRUARY 2004.

[5] P. Marcuello, A.Gonzalez and J. Tubella. Thread Partitioning and Value Prediction for Exploiting Speculative Thread-Level Parallelism. IEEE TRANSACTIONS ON COMPUTERS, VOL.53, NO.2, pp.114-125, FEBRUARY 2004.

[6] D. Carney, U. Cetintemel, M. Cherniack, C. Convey, S. Lee, G. Seidman, M. Stonebraker, N. Tatbul, and S. B. Zdonik. Monitoring Streams - A New Class of Data Management Applications. In VLDB 2002, Proceedings of 28th International Conference on Very Large Data Bases, pp.215-226, August, 2002, Hong Kong, China, 2002.