データマイニングによる時系列からの特徴抽出

2C－5

佐藤 嘉則, 前田 章
（株）日立製作所 システム開発研究所
e-mail: y-satou@sdl.hitachi.co.jp, maeda@sdl.hitachi.co.jp

1. はじめに

これまで各種プラントでは知識情報処理技術を応用したシステムの導入が進められてきたが、知識獲得ボトルネックが保守管理、システム更新の大きな障害になる場合が多い。一方プラントにおける計算機環境の発達により、詳細で大量のデータを取得することが可能になってきた。そこで本報告では時系列記号化技術[4]、association rule 方式[5]を用い、データマイニングによりプラントデータの時系列データから特徴を抽出する方法を提案する。

2. データマイニングの枠組み

データマイニングはデータベースに蓄積された多変量、かつ大量データ中の規則性を抽出を目的とし、機械学習の知識獲得技術が背景となっており[1]、手法としては組指向、属性指向のものが提案されている[2]。

これまでデータマイニングの適用分野としては顧客情報分析、プラント操業データ分析等があり、それまで知られていなかった妥当なルールを抽出することに成功している[3]。

しかしこれらの適用では多変量空間の分析に主眼が置かれており、データの表現形式、ルール抽出方式に起因する制約のため、時間的因果関係を扱うには問題があった。本研究ではこれらの問題点を念頭に置き、時系列データから時間的因果関係を抽出する方法を提案する。

3. 時系列データからの特徴抽出

時系列データマイニングでは、１個以上の時系列を生じる変数が与えられたとき、任意の変数値の組 X に着目し、結果 Y が強い関連性をもって生じる場合、これをルール $X \Rightarrow Y$ として抽出する。X をルール条件部、Y をルール結論部と呼ぶ。

3.1. 時系列を扱う際に生じる問題点

しかし時系列を扱う場合、以下の点が問題になる。

- サンプリング周期にもよるが、同じ値が時間的に連続して生じる場合があり、データが冗長であり、データ量が膨大になる。
- 時系列データの場合、瞬間的な値だけでなく、データの変動パターンに意味がある。変動パターンを統計量算出の単位とした場合、パターンマッチングの問題が生じる。

4. アプローチ

そこで時系列データの変動パターンを記号値に変換し[4]、これらの記号値に基づいて統計量を算出する。記号化によりデータの冗長性も同時に除外される。

また、時系列データの性質を考慮して、同じ時刻に生じた各変数値の変動パターンの組に基づいて統計量の算出を行うこととする。ここでは変動パターンの組を事例と呼ぶ。ただし、変動パターン自体が時間的な幅をもち、また変数間に遅れ時間が存在することを考慮して、事例を規定する時刻には幅を持たせる。

例えば、二つの変数 P_1, P_2 がある場合、「時刻 $[T-\Delta T, T]$ に生じた P_1 が上昇、P_2 が下降」が事例となる。

4.1. アルゴリズム

時系列データは時間順に並んで格納されていることが一般的であり、記号化したデータを発生時刻、記号値のペアをレコードとして格納した場合、１個の事例には１個以上の連続したレコードが含まれる。

このようなデータ構造を対象としてルール抽出を行う手法としては、association rule 方式がある[5]。association rule 方式では POS データを対象とするため、ルール抽出は教師なし学習の範疇で実現される。association rule 方式ではまず、変数値組の生起確率を計算し、これが外部から与えたしきい値以上のものを対象としてルール結論部となる変数値を探しに行く。

ところがプラントで時系列データを分析する場合、システムの異常や、オペレータの操作情報等何らかの教師情報を得られることが多い。そこで本報告で

はシステム外部から与える定義に基づいて事例を正例と負例に分け、教師あり学習に基づいてルール抽出を行う。例えば、システム異常発生時刻の近傍で生じた変数値の組を正例とし、正例集合に特徴的なルールを負例集合との比較により抽出する。ここで負例は正例とは時間的に disjoint な集合のことを指す。

実際の計算では association rule 方式により変数値組の生起確率を算出する際、以下の評価関数[6]に基づいてルールそのものの算出を行う。

$$u = P(X)^{\beta} P(Y|X) \log \frac{P(Y|X)}{P(Y)}$$

ただし、X は変数値の組、Y は事例が正例に属する状態を表す。負例との比較により評価値が正になるものをルールとして採択する。

5. 実験

ファジィ制御されている倒立振子の観測データからの制御ロジックの再現実験を行った（図 1）。観測される変数は台車位置 X, 台車速度 $\dot{X}$、振り子角度 θ、振り子角速度 $\dot{\theta}$ でありそれぞれ 10%の観測誤差が含まれているものと、一様乱数でかき混ぜた意味のない変数を 4 つ与えた。ただし単純化のため変数の記号化は時系列記号化ではなく、変数値のレンジを 5 分割し、分割された領域を小さい順に negative, little negative, almost zero , little positive, positive として記号値を与えた。台車に与えた力 F が正の場合と負の場合とで事例を分け、負の場合を正例としたときのルール生成を行ったものの結果を表 1に示す。ルール 3、ルール 5 は正しい制御ロジックを示しており、ほかのルールも物理現象として正しいものになっている。

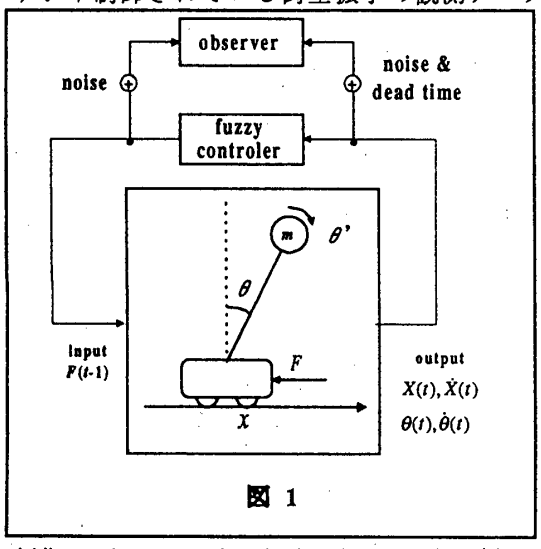

図 1

No.	condition	P(X)	P(X|Y)	u
0	X'=a little negative	0.486	1	0.0183
1	Θ'=a little positive	0.408	1	0.0159
2	Θ'=positive	0.338	1	0.0137
3	Θ=negative	0.29	1	0.0121
4	X'=negative	0.256	1	0.0109
5	Θ=a little negative	0.252	0.992	0.00809
6	X=positive	0.148	0.986	0.00405

表 1：F が負の時のルール

6. まとめ

本報告では時系列記号化により変動パターンの記号値を生成し、教師あり学習方式を用いて時系列から特徴抽出を行う手法を提案した。本手法では可読性が高い特徴的なルールを抽出することが可能であり、大規模、複雑な分析対象に対しても適用することが容易である。

7. 参考文献

[1] G.Pietetsky-Shapiro and W.J. Frawley, editors. *Knowledge Discovery in Databases*. AAAI press / The MIT press, Menlo Park, CA, 1991.

[2]川野浩之, 西尾章治郎, Han. J.:データベースからの知識獲得技術, 人工知能学会誌, vol.10, no.1, pp. 38-44, 1995

[3]下田睦、前田章、服部哲、斎藤裕、データマイニング技術を用いた鉄鋼圧延プラント異常診断システムの開発、電気学会産業応用シンポジウム予稿集 (1995)

[4] Takeshi Nishiya, *A Signal-to-Symbol Transformation Method of Time-Series Data for Detecting Signs of a Process Change*, Proc. of IEEE International Workshop on NEURO-FUZZY CONTROL, pages 345-351, Murolan, Japan, March 22-23,1993

[5] Rakesh Agrawal,Ramakrishnan Srikant, *Fast Algorithms for Mining Association Rules*, Proceeding of the 20th VLDB Conference Santiago, Chile, 1994

[6]芦田仁史、前田章、高橋ヨリ:データマイニングによる特徴的ルール生成方式、情報処理全国大会講演論文集、vol. 3, pp. 19-20, 1994 年 3 月