学習分類子システムを用いた プロセス時系列からのデータマイニング

倉橋　節也*1　　寺野　隆雄*2

*1　東京電機産業株式会社　*2　筑波大学経営システム科学専攻
kura@tokyo-densan.co.jp　terano@gssm.otsuka.tsukuba.ac.jp

本研究は，連続プラントを対象とした大量の時系列データから，プロセス応答モデルを構築する．また現在の操業データからの将来データの予測するプロセス予測モデルと，運転員に対するプラントの操作ガイダンスを行うプロセス制御ルールの発見的探索手法を提案する．そして，実際のバイオプラントの操業データに適用した結果を報告する．モデルの基本的な考え方は，時系列データ間の相互相関係数最大化，ニューラルネット時系列予測，MDL規準と改善率にもとづくアソシエーションルールの分類子学習である．本解析手法を用いて，バイオプラントにおける実プロセスデータを解析し，プロセス応答モデルの予測性能の有効性，およびC4.5による決定木分析との比較からMDL規準と改善率に基づく分類子学習の有効性を実証する．

Data Mining from Plant Process Time Series by a Learning Classifier System

Setsuya Kurahashi*1,　Takao Terano*2
*1 Tokyo Denki Sangyo Co., Ltd.
*2 Graduate School of Business Sciences, University of Tsukuba

Continuation processes in chemical and/or biotechnical plants always generate a large amount of time series data. However, since conventional process models are described as a set of control models, it is difficult to explain the complicated and active plant behaviors. Based on the background, this research proposes a novel method to develop a process response model from continuous time-series data. The method consists of the following phases: 1) Collect continuous process data at each tag point in a target plant; 2) Normalize the data in the interval between zero and one; 3) Get the delay time, which maximizes the correlation between given two time series data; 4) Select tags with the higher correlation; 5) Develop a process response model to describe the relations among the process data using the delay time and the correlation values; 6) Develop a process prediction model via several tag points data using a neural network; 7) Discover control rules from the process prediction model using Learning Classifier system. The main contribution of the research is to establish a method to mine a set of meaningful control rules from Learning Classifier System using the Minimal Description Length criteria.

1．はじめに

化学プラントやバイオプラントといった連続プラントにおいて時系列データの解析が再び注目されている．それは，ネットワークや情報システムの低コスト化により，プロセスデータを大量に蓄えるプラント情報システム（PIMS）が導入しやすくなったことによる．

従来のプロセスモデルは，個々の反応工程を一時遅れ関数などの伝達関数によって記述し，その集合としてモデルを構築してきた．しかし，投入原料の成分変動や運転条件の違いなどの組み合わせによって，プロセスの状態は大きく変化してしまう．PIMS によって次々と蓄えられる日々の実データからプロセス特性を得ることができれば，これらの問題は解決する．このような時系列データから妥当性の高い知識を獲得することはアクティブマイニングの重要なテーマの１つである。

プロセスの状態を表す従来のモデルは，単純化し過ぎていて現実のプラント操作には不十分であることが多かった．これは，モデルが現実のデータに基づかず，理想的な反応モデルの集合として構築されていることに起因していた．また，実際のプロセスデータに基づいたモデルであっても，プロセス応答特性をブラックボックスモデルとして構築している場合，オペレータが理解できないような複雑な結果のみを示してしまうことがあった．

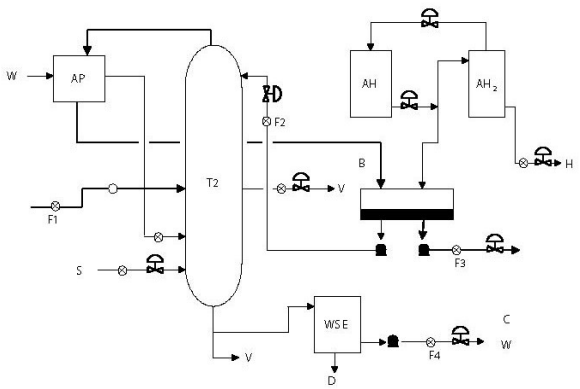
本研究は，連続プラントを対象とした大量の時系列データから，プロセス応答モデルを構築する．また現在の操業データから将来データを予測するプロセス予

測モデルと，運転員に対するプラントの操作ガイダンスを行うプロセス制御ルールの発見的探索手法を提案する．そして，実際のバイオプラントの操業データに適用した結果を報告する．モデルの基本的な考え方は，時系列データ間の相互相関係数最大化，ニューラルネット時系列予測，MDL(Minimum Description Length)基準\cite{Rissanen:78}と改善率にもとづくアソシエーションルールの分類子学習である．

本解析手法を用いて，バイオプラントにおける実プロセスデータを解析し，プロセス応答モデルの予測性能の有効性，およびC4.5による決定木分析との比較からMDL規準と改善率に基づく分類子学習の有効性を実証する．

２．対象とするバイオプラント

対象とするプロセスはFig1.に示すような蒸留塔を中心とするバイオプラントの反応工程である．原料が蒸留塔に注入され低圧処理の結果成分分離が行われる．実際はこのような蒸留塔や精留塔を中心とした設備が複数カスケードに存在する．本研究ではその中のひとつの設備を対象に分析を行った．



Fi.1:　対象としたバイオプラント

プロセスデータは一般に温度℃，圧力（Pascal），流量(Nm^3/h)など様々な工業単位をもっている．そのためこれらを正規化する必要がある．　に正規化されたデータのトレンドグラフを示す．しかしこのように，正規化されたデータを見ただけでは，どのような関連があるのかを読み取ることは不可能に近い．実際の操業者はこのようなデータを見るのではなく．過去の経験に基づく個々のループ制御操作によって全体をコントロールしている．本研究では，この複雑に見える時系列データから，意味のある情報を取り出すことを目的とする．

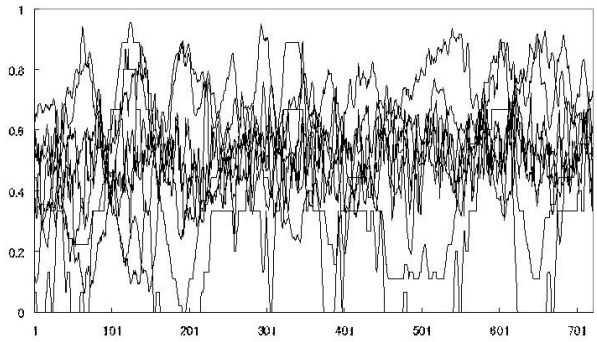


Fig.　2:　正規化したプロセスデータ

３．制御ルールの発見的探索

プロセスデータは，時系列であること，相互相関を持つこと，応答特性が異なること，といった特徴をもっている．その中で，当該プラントの適切な制御ルールを学習することは，担当するオペレータにとって重要な目標となっている．しかしプラントをモデル化するには，多変数予測モデルなどの高価なパッケージソフトを用いたにしても，複雑なモデル化作業が必要となるため，実際の運転操作において，いつ，どの制御ループを，どれだけ操作するかといった判断は，そのほとんどを経験と勘に頼っている．大量に蓄積されたプロセスデータから制御ルールを発見する手法として決定木などによる探索も可能だが，大規模な枝が発生してしまうことが多い．　Fig.3に決定木分析手法として標準的なC4.5を適用した結果の一部を示す．

途中を省略してあるが，枝刈りを行ってもこのような大規模な木が出現し，オペレータに分かりやすく意味のあるガイダンスを与えることはできない．また，顧客行動分析などに比べ，プロセスモデルの場合より確実な制御応答性能を求められる．そのため，簡潔なモデルで，確実な，意味のある情報を導出することが必要となる．

本章ではこれを解決するために基準と改善率に基づくクラシファイアシステムを提案する．　基準はモデルの複雑さとデータの複雑さを最小にするものとして提案され研究が行われている．その意味で簡潔で確実な結果を得るには都合が良い．しかし，当たり前な結果を導き出すことに対して考慮されているものではない．当たり前とは，頻

繁に出現する事例で，経験的に判断できるものである．そこで，バスケット分析で用いられる改善率を考慮したモデルを提案する．改善率とは対象とする事象の全事象の中のシェアと，ある条件によって抽出された部分事象の中でその事象が出現する頻度を表す信頼度との比である．

クラシファイアシステムとしてビッツアプローチを採用する．学習は全データを対象とせず，ランダムサンプリングによって事象の分布を推定する方法を用い，対象とする事象が多いことへ対応する．分類子の構造を以下に示す．

p = (p^1_1 ^ p^1_2^…^ p^1_m)v=(p^2_1 ^ p^2_2^…^ p^2_m)v…

r = r1, r2, …, rn

```
C4.5 [release 5] decision tree generator
----------------------------------------
Options:
Pruning confidence level 25%
Read 642 cases (8 attributes)

Simplified Decision Tree:

F3L <= 1 :
|   F3L <= 0 : 3 (40.0/2.6)
|   F3L > 0 :
|   |   F2L > 2 : 1 (8.0/1.3)
|   |   F2L <= 2 :
|   |   |   F1L <= 1 : 2 (5.0/2.3)
|   |   |   F1L > 1 :
|   |   |   |   F4L <= 2 :
|   |   |   |   |   F2L <= 1 :
|   |   |   |   |   |   F1L <= 2 :
|   |   |   |   |   |   |   F4L > 1 : 3 (29.0/12.3)
|   |   |   |   |   |   |   F4L <= 1 :
|   |   |   |   |   |   |   |   F2L <= 0 : 2 (8.0/4.5)
|   |   |   |   |   |   |   |   F2L > 0 :
|   |   |   |   |   |   |   |   |   F3DX <= 1 : 1 (10.0/5.6)
|   |   |   |   |   |   |   |   |   F3DX > 1 : 3 (4.0/3.1)

（途中省略）

|   |   |   |   |   |   |   |   F4L <= 1 :
|   |   |   |   |   |   |   |   |   F4DX <= 1 : 1 (9.0/3.5)
|   |   |   |   |   |   |   |   |   F4DX > 1 : 0 (3.0/2.1)
|   |   |   |   |   F4L > 2 :
|   |   |   |   |   |   F2DX <= 1 : 0 (3.0/2.1)
|   |   |   |   |   |   F2DX > 1 : 2 (2.0/1.0)

Tree saved

Evaluation on training data (642 items):

            Before Pruning           After Pruning
            ----------------   ---------------------------
            Size      Errors   Size      Errors    Estimate

             185  118(18.4%)      93  144(22.4%)    (33.7%)
```

Fig.3: 決定木分析の結果

ここで，**p**は前件部の条件式を表し，**r**は後件部の結果を表す．ただし，結果はその事象のとりうる全種類のヒット数をカウントする．これにより，前件部に一致した事象の後件部事象の信頼度を得ることができる．バスケット分析における改善率は，一般に以下の式で表される．

$$improvement = P(ri) / P(ri \mid p)$$

ここで，$P(ri)$は条件なしで**r***i*という後件部が出現する確率，$P(ri \mid p)$は前件部**p**の条件付きで**r***i*という後件部が出現する確率を表わす．

これは，絞込みを行う前の情報量と，ある条件によって絞込みを行った後の情報量との差と考えることができる．そこで初期状態でのシェアをMDL原理の記述長として表現し，クラシファイアシステムによって分類されたデータ長とモデル長との差を最大とするように，MDL規準を拡張する．この結果，改善される情報量とMDL原理における記述長を同時に扱うことが可能となる．学習した分類子のウェイトの算出方法を示す．

$$DL_{first} = dataLength_f + \mathrm{mod}\, elLength_f$$

$$DL_{last} = dataLength_l + \mathrm{mod}\, elLength_l$$

$$weight = DL_{first} - DL_{last}$$

ここで$dataLength_f, \mathrm{mod}\, elLength_f$は初期状態でのデータ記述長とモデル記述長を示し，$dataLength_l$と$modelLength_l$は最終状態でのデータ記述長とモデル記述長を表わす．前件部にヒットする分類子を全て評価し，その分類結果をカウントすることで，分類誤差の推定値を計算する．そして，判定結果の分類の中で最小の$dataLength$を選択してMDL規準によるウェイトを求める．このウェイトを最大化するようにクライファイアを実行する．以下にMDL規準を示す．

$$\min_k DL_k = dataLength_k + modelLength_k$$

$$dataLength = mH(\frac{m_1}{m}) + \frac{1}{2}\log(\frac{mp}{2}) + o(1)$$

$$modelLength = \sum_{i=1}^{k} c_i(1 + \log t_i)$$

$$H(x) = -p\log(x) - (1-x)\log(1-x)$$

c_i : 前件部の条件が#の時0，それ以外は1

t_i : 各条件のデータ分割数

4．モデル構築

プロセスデータ解析のためのモデル構築を，実際のデータを用いて行った例を説明する．基本的な構築の手順を次に示す．

(1)データの正規化：さまざまな工業単位を持ったプロセスデータを同一に扱うために正規化する．

(2)相互相関分析：二つのプロセスデータを選択し，時間を徐々にシフトして最大の相関を示す時間差を探索す

る．

(3)解析対象とするプロセスデータの選択：実プラントの配管系統図などを参考に，解析の対象とする相互相関係数の高いプロセスデータを選択する．

(4)プロセス応答モデル：シフト時間と相関係数から，プロセスデータ間の関係を応答モデルとして記述する．

(5)プロセス予測モデル：プロセス応答モデルから，注目するプロセスデータを従属変数とした複数のプロセスデータを選択し，ニューラルネットなどによって予測モデルを構築する．

(6)制御ルールの発見：プロセス応答モデルから，注目するプロセスデータをクラスとして，クラシファイアシステムを実行し，制御ルールを見出す．

４．１　応答モデル

多変数の場合は主成分分析などの統計手法が一般的だが，時系列解析には自己回帰モデルが用いられることが多い．自己回帰モデルの多変数への拡張も行われているが，連続プロセスデータの場合，時系列データであってもそれぞれのデータ間に広く相関が認められる．そこで，ここではデータ間の関係に着目し正規化された時系列データの相互相関係数を以下の操作により求める．

(1)対となるプロセスタグの時系列データx,yを選択する．

(2)kをそれぞれのタグの時間シフト量とし，k次相互相関数r_{xy}(k)を次式にて求める．

$$r_{xy}(k)=\frac{\sum_{t=k+1}^{T}(x_{t-k}-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=k+1}^{T}(x_{t-k}-\bar{x})^2}\sqrt{\sum_{t=1}^{T}(y_t-\bar{y})^2}}$$

(3)最大の相関を示すkを求める．

$$\max_k r_{xy}(k)$$

Fig.４にkを前後60分シフトしたk次相互相関係数の例を示す．縦軸が相関係数、横軸がシフト時間を表す．ひとつのプロセスデータを基準として，他の時系列データをシフトした時の相互相関係数の変化が分かる．Fig.5は，相関がほとんど見られない例を示している．

この操作を全てのデータの組み合わせで実行して得られるのが，最大相関係数表とシフト時間表である．その一部をTable1,2に示す．

この表により，相関の高いタグを抽出することができる．例えば±0.4以上の相関係数を持つタグは，F4とF5およびF3とF4の対になる．その時のシフト時間は５分と１０分である．このようにして多数の時系列タグデータから相関の高いタグを抽出し，プロセス応答モデルを構築することができる． Fig.6にその例を示す．

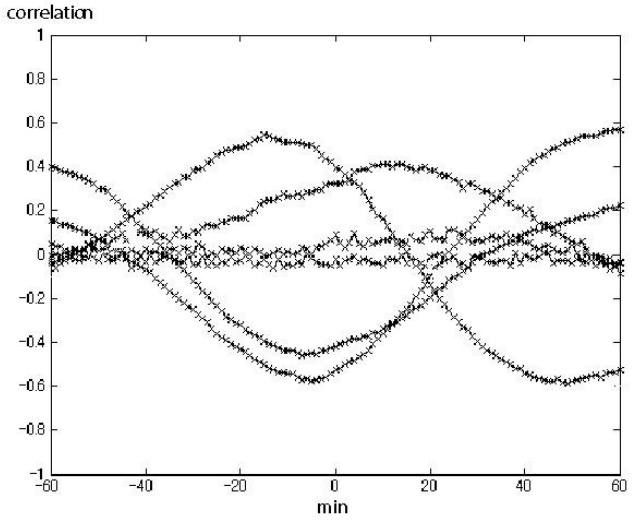


Fig.4:　k次相互相関係数グラフ１

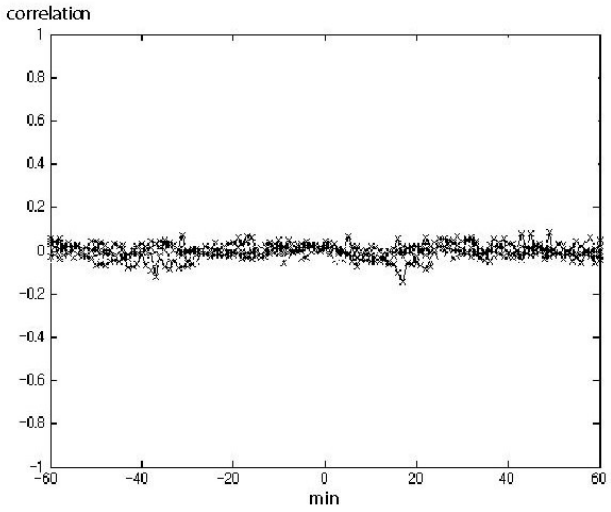


Fig.5:　k次相互相関係数グラフ２

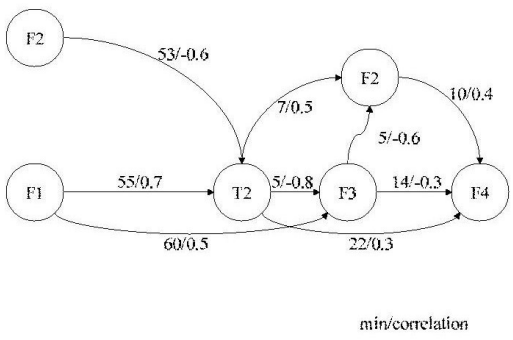


Fig.6:　プロセス応答モデル

これによれば，原料の流入流量を表すF1タグと蒸留塔中央温度を表すT2タグには55分のシフトで0.7の正の相関が認められる．つまりの流量が増加すれば55分後にT1の温度が上昇することを示している．同じようにT2の温度が上昇すれば5分後にF3の流量が減少し，その14分後にF4の流量が増加する．ここで注目すべきなのはF2のタグである．これはT2に負の相関を示すと同時に同じT2とF4に正の相関を示している．理由はプラントの配管にある．蒸留塔から抽出分離された液はバッファタンクに蓄えられた後，その一部が同じ蒸留塔に戻るような配管になっている．この戻り配管によって， F2の流量が増加するとT2温度が53分後に低下し，その7分後にF2自身の流量が増加することになる．プロセスタグ間の時系列相関と時間シフトの情報により，物理的なプラント配管系統図からだけでは知ることができないプロセス応答の構造を簡潔に示すことができる．

Table 1: 最大相関係数

最大相関	F1	F2	F3	F4	F5
F1	1.00	-0.12	-0.14	-0.06	0.04
F2	-0.12	1.00	-0.12	0.11	0.11
F3	-0.14	-0.12	1.00	0.41	-0.32
F4	-0.06	0.11	0.41	1.00	-0.57
F5	0.04	0.11	-0.32	-0.57	1.00

Table 2: シフト時間のまとめ

最大相関	F1	F2	F3	F4	F5
F1	0	-37	17	13	-60
F2	37	0	34	-25	-60
F3	-17	-34	0	-10	-14
F4	-13	25	10	0	-5
F5	60	60	14	5	0

4．2予測モデル

プロセス応答モデルで求めたタグデータを用いて，プロセス予測モデルを構築する．品質に大きな影響を与える温度T2を予測するものである．ここでは非線形なデータを扱うためニューラルネットでモデルを構築する．誤差逆伝播学習アルゴリズムで学習を行う．

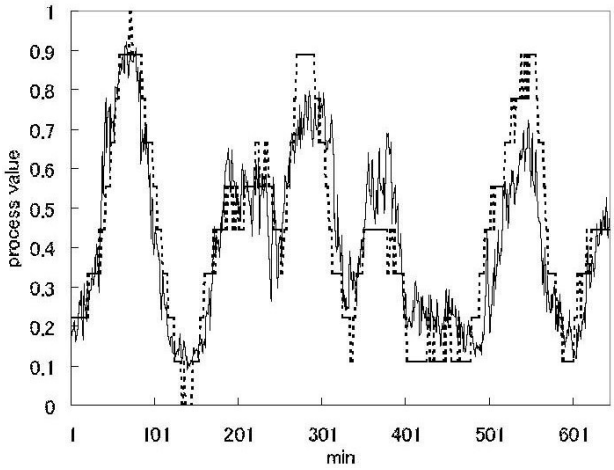


Fig.7: プロセス予測モデル

Fig.7に2万回学習した結果を示す．点線が実現値，実線が本プロセス予測モデルの予測値である．学習したモデルによる予測値と実現値との比較を決定係数R^2を求めると，自由度修正済み決定係数0.74となっており，比較的当てはまりがよい．ここで用いたプロセスタグデータの関係は，T2＝f(F1,F2)のようになっている． F1，F2の応答特性はそれぞれT2に対して55分前で相関係数が0.7， 53分前で相関係数-0.6がとなっていた．このことは，F1，F2のプロセスデータを測定すれば，少なくとも53分後のT2の温度が予測できることを表している．これは，プラントオペレータにとって重要な情報となる．

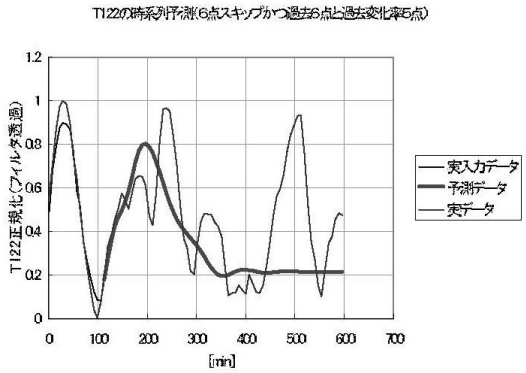


Fig.8: 自己相関モデル

次に，自己相関モデルを同じくニューラルネットワークでモデル化し，さらに未来の予測値を求めてみた結果をFig.8に示す．過去のデータ6点と変化率5点を用いて計算を行った． 50分程度までの予測はおおよそ実データと一致しているが，それ以上は外れてきている．上流工程の変

化が大きく反映するプロセスであるため，自己相関モデルの信頼性は高くはない．ARMAモデルにおける偏自己相関係数による分析結果は1/6サンプリングでAR(2)となっていた．

４．３　制御ルールの発見的探索

上流工程のプロセスデータから下流工程の状態を予測することができたが，実際の操業ではどのように下流工程の品質を安定させるかが重要な目標となる．ニューラルネットモデルは高いフィッティングを与えてくれるが，基本的にブラックボックスモデルとなるため，運転オペレータに対する分かりやすい制御方針を示してくれるものではない．そこで，プロセス応答モデルで得られた相関の高いタグのデータを対象にMDL規準と改善率に基づくクラシファイアシステムによる制御ルール探索を行う．得られた分類子の例を以下に示す．このときの改善率は3.1，MDL値は32.9ビットとなっている．また，改善率を考慮した場合の分類子を続けて示す．このときの改善率は6.6，値は54.8ビットとなっている．前者はより簡潔なモデルとなっているが，改善率が前者よりも低く「当たり前な結果」に近い．

```
MDL:
  75% < F2 and 75% < F3
    then 75% < T2
MDL+Improvement:
  25% < F3 <= 50% and 75% < F4 and F3 is down
    then 75% < T2
```

また，別のプロセスデータの場合を以下に示す．このときの改善率は3.9，MDL値は121.3ビットとなっている．また，改善率を考慮した場合の分類子を続けて示す．このときの改善率は5.1，MDL値は121.3ビットとなっている．この場合もMDL規準に加えて改善率を考慮した場合に改善率が高く，より意外性のあるルール発見となっている．

```
MDL:
  (50% < F1 <= 75% and 25% < F4 <= 50%
   and F4 is up) or 75% < F2
    then T2 <= 25%
MDL+Improvement:
  (75%<F2 and 50%<F3<= 75% and F4 is up)
   or 75% < F2
    then T2 <= 25%
```

５．おわりに

本論文では，大量の時系列プロセスデータから最大時系列相関の方法を用いてプロセス応答モデルを構築しタグ間の関係を遅れ時間と相関係数によって示すことを行った．そしてニューラルネットによるプロセス予測モデルを作成し，現在のプロセスデータから未来の状態を予測できることを実証した．そして，オペレータに対する運転支援として，簡潔で信頼性が高く意味のある操作を指示することのできるMDL規準と改善率に基づくクラシファイアシステムを提案し，実際のプラントデータを用いて制御ルールを見出した．今後，作業員の手動操作などのイベント情報も同時に扱えるモデルへ拡張を行う予定である．

参考文献

[1] C. Adami: Complexity of Simple Living Systems, Introduction to Artificial Life, Springer-Verlag NY,113/138 (1998)

[2] E. Suzuki and S. Tsumoto: Evaluating Hypothesis-Driven Exception-Rule Discovery with Medical Data Sets, Knowledge Discovery and Data Mining(ed; Terano, T., Liu, H., Chen, A.L.P.), 208/211 (2000)

[3] P. Adriaans and D. Zantinge: Data Mining, Addison-Wesley (1996)

[4] J. Rissanen: Modeling by shortest data description, Automatica, Vol. 14, 465/471 (1978)

[5] 山西健司: 統計的モデル選択と機械学習，計測と制御，Vol. 38, No. 7, 420/426 (1999)

[6] 和田卓也，堀内匠，元田浩，鷲尾隆: Ripple Down Rules法における知識獲得と帰納学習の統合化の試み，MYCOM, 1, 66/73 (2000)

[7] D. J. Berndt and J. Cliford: Finding Patterns in Time Series: A Dynamic Programming Approach, Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors.), MIT Press, Cambridge, MA, 229/248 (1996)