計測自動制御学会東北支部 第198回研究集会 (2001.11.17)
資料番号 198-4

プラント運転支援のためのＫＤＤＭツールボックスの試作とその応用

A trial production and application of the KDDM toolbox for plant operation support

○望月禎仁 , 山下善之 , 鈴木睦

○Yoshihito Mochiduki , Yoshiyuki Yamashita and Mutsumi Suzuki

東北大学大学院工学研究科化学工学専攻

Depatrment of Chemical Engineering, Tohoku University

キーワード： 知識発見(Knowledge Discovery) データマイニング (Data Mining), 機械学習(Machine Learning), プラントモニタリング(Plant monitoring) 異常検出(fault detection)

連絡先： 〒980-8579 仙台市青葉区新巻字青葉07 東北大学大学院 工学研究科 化学工学専攻 鈴木研究室
望月禎仁, Tel.: (022)217-7268, Fax.: (022)217-7293, E-mail: motizuki@pse.che.tohoku.ac.jp

1. 緒言

現在の化学プラントではDCS(分散型制御システム)の導入により、大量のプラント運転データが保存されており、蓄積されているデータの系統的な利用が望まれている[1, 2]。このデータの中にはプラントの設備・運転・制御に関する有用な情報が含まれているはずであり、現在KDD(Knowledge Discovery in Database)を利用したデータ解析の研究が行なわれている。しかし、蓄積されたデータに対するデータ解析の方法は対象や目的に応じて変わってくるため複数のKDD手法によるデータ解析を行う必要がある。そこで、様々な手法をツールボックス的にまとめて利用できる統合的なデータ解析を行なう環境を提案する。

今回はツールボックスに関する構想とツールボックスの試作を行ない、応用としてTennessee Eastman プラントシミュレーターから得られた異常発生時の時系列データに対して代表的なKDD手法の一例を適用しプラントの状態検出を試みた。また、その結果と従来から状態検出に用いられている管理図(Shewhart Chart)、累積和管理図(Cusum Chart)の2種類との比較を行なった。

2. KDDMツールボックス

KDDプロセスのデータ収集、前処理、データマイニングアルゴリズム実行の適用を統合的に実施するためのツールボックスの試作を行なった。ツールボックスの構造をFig.1に示す。

実際にはまだ構想段階の部分もあり使用できるツールも多くないため、今回はデータマイニングツールを一つに絞ってデータ解析を行なった。

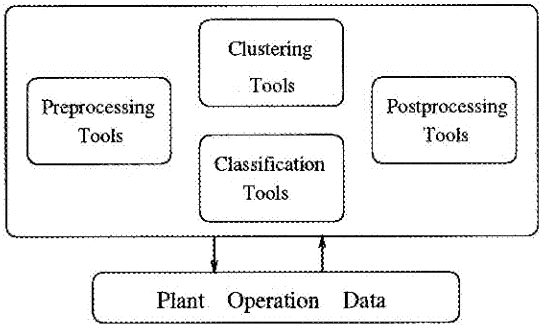

Fig. 1 KDDM Tool Box

2.1 データ入出力

ツールボックスの構築にはフリーで配布されている数値計算用プログラミング言語MaTX[3]を主に使用した。データの入出力もこのMaTXに準拠している。それぞれのアルゴリズムのインプリメントはC言語・Matlab・Java等、色々な言語で行なわれている。

2.2 前処理

データのノイズ除去や定性的な情報を与えるための前処理の手法をいくつか集めた部分である。代表的なものに、微係数・時間窓・FFT・各種フィルター・PCAなどがある。

今回は、各属性変数の一次微分をデータ中に加えることにより経時変化の情報を付加した後、フィルターを用いてノイズの平滑化を行なった。

2.3 データマイニングツール

データマイニングツールはKDDの中で実際にデータから情報を抽出するものでありKDDの中心的な役割を果たす。データマイニング手法には様々な方法があるが主なものとしてClassification型とClustering型の2つが挙げられる。

Classificationとは得られたデータを予め定めてあるいくつかのクラスに分類することを言う。代表的な手法として、決定木に基づく方法(C4.5,CART等)、教師付ニューラルネット等が挙げられる。

Clusteringとはデータを記述する有限個のカテゴリーを同定することを言う。代表的な手法として、教師無しニューラルネット(ART2,SOM等)、統計的な手法(Autoclass等)が挙げられる。

上記の他にもデータマイニングツールとして知られているものは多くあるが、今回は他のツールと比較して計算時間が短く、性能も優れていることから様々な分野で実用に用いられているC4.5アルゴリズム[4]を使用することとした。以下にC4.5の説明を行なう。

2.4 C4.5

C4.5とは決定木(Decision tree)と呼ばれる分類器(Fig.2)に基づいた機械学習を行なうClassification型のデータマイニングアルゴリズムである。

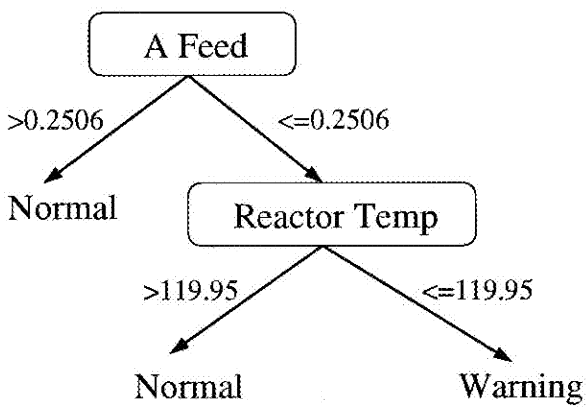

Fig. 2 Decision Tree

J.R.Quinlanが作成したID3アルゴリズムをQuinlan自身が改良しC4.5として完成させた。以下でC4.5の決定木作成基準について説明する。

C4.5では決定木を作成する基準として情報エントロピーを用いる。データTの中に存在するクラスが2種類(例:正常・異常、賛成・反対 等)の時、2つのクラスをP,Nとし、そのクラスに属するデータの個数をp,nとする。そのときクラスを同定するのに必要な平均情報量$info(T)$は

$$info(T) = -\frac{p}{(p+n)} log_2 \frac{p}{(p+n)} - \frac{n}{(p+n)} log_2 \frac{n}{(p+n)} \quad (1)$$

で表される。この$info(T)$を集合Tの情報エントロピーという。ここで、集合T中にあるいずれかの属性で集合Tの分割を行なうがこの分割をテストXといい、このテストXにより集合Tが部分集合$T_1, T_2, \cdots T_n$に細かく分割されたとき、分割前と同様の情報量の評価を計算すると

$$info_x(T) = -\sum_{i=1}^{n} \frac{p}{(p+n)} log_2 \frac{p}{(p+n)} \qquad (2)$$

となる、そしてこれらの差である

$$gain(X) = info(T) - info_x(T) \qquad (3)$$

は、テストXで集合Tを分割することで得られる情報量を表わすことになり$gain(X)$自体は情報量利得と呼ばれる。C4.5の旧版であるID3ではこの情報量利得が最大になるようテストXを選択して決定木を成長させて行く。この基準$gain(X)$のことを利得基準と呼ぶ。

しかしこの基準では部分集合$T_1, T_2, \cdots T_n$のnが非常に大きくなる変数を偏重する欠陥があるため、C4.5では分割自体に必要な情報量$splitinfo(X)$で利得基準$gain(X)$の規格化を行なう。$splitinfo(X)$は次の式で表わされる。

$$splitinfo(X) = -\sum_{i=1}^{n} \frac{p_i + n_i}{p+n} log_2 \frac{p_i + n_i}{p+n} \qquad (4)$$

この$splitinfo(X)$で$gain(X)$を規格化した値は利得比$gainratio(X)$と呼ばれ、C4.5では利得基準$gain(X)$が平均以上という条件下で最も$gainratio(X)$が大きくなるテストを決定木を成長させるテストとして選択し決定木の先頭に配置する。

この作業を再帰的に繰り返し、情報量利得が得られなくなるまで決定木を成長させて行くこととなる。

2.4.1 C4.5rules

前節で得られた決定木が非常に大きく複雑になり人間に理解が難しくなった場合のためにC4.5にはC4.5rulesという決定木をIf-Thenルールに置き換えるC4.5rules機能が付属している。If-Thenルールの例をTable.1に示す。このIf-thenルールへの変換

Table 1 If-Then rule

If	(A feed $\geqq$ 0.2506)
and	(Reactor temp $\leqq$ 119.95)
Then	(Class = Warning)

により人間が解釈しやすくデータの特徴の把握も容易になる。

If-Thenルールへ変換する場合、ただ単純に決定木をIf-Thenルール化するのでは無く、決定木中に現れた不要な条件などを取り除くことで得られるIf-Thenルールの簡略化を行なっている。

2.4.2 性能評価

得られた決定木・If-Thenルールに対して決定木を作成したときとは別のデータ（テストデータ）を適用することで、決定木・If-Thenルールの性能評価を行なうことができる。この場合、テストデータが元々持っているクラスと、テストデータを決定木・If-Thenルールによって分類した結果のクラスを比較し、その誤り率で評価を行なう。

2.5 CUSUM

工業的に応用されている統計的方法の代表的なものに管理図(Shewhart Chart)がある。今回はShewhart Chartと、その改良型である累積和管理図(CUSUM) の2種類をC4.5で得られたルールを用いた状態診断と比較し検討を行なった。以下にCUSUMの説明を行なう。

まずCUSUMではプロセスの時系列観測データxに設定値μ_{TR}を設ける。また、設定値の上側の閾値k_Hと下側の閾値k_Lを設定し、観測データがこの閾値を越えた分($X(t)$の太線部分)についてS_H,S_Lの累積計算を行なう(Fig.3)。

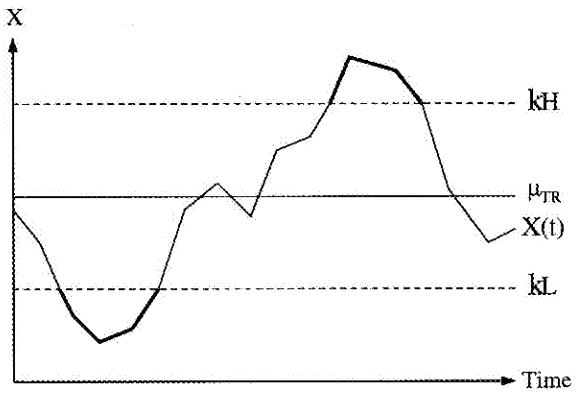

Fig. 3　An Example Process Trend

S_H,S_Lの式を以下に示す。

$$S_H(t) = max[0, S_H(t-1) + x(t) - \mu_{TR} - k_H] \quad (5)$$

$$S_L(t) = max[0, S_L(t-1) - x(t) + \mu_{TR} - k_L] \quad (6)$$

$$S_H(0) = 0, S_L(0) = 0 \quad (7)$$

上側の閾値k_Hを越えたときには(1)によりS_Hが累積計算され、下側の閾値k_Lを越えたときには(2)によりS_Lが累積計算される。最終的には(4)により$CUSUM(t)$が計算される。

$$CUSUM(t) = max[\|S_H\|, \|S_L\|] \quad (8)$$

そして$CUSUM(t)$がある閾値Hを越えた場合にプロセスの運転状態の変化・異常などが発生したと判断する(Fig.4)。

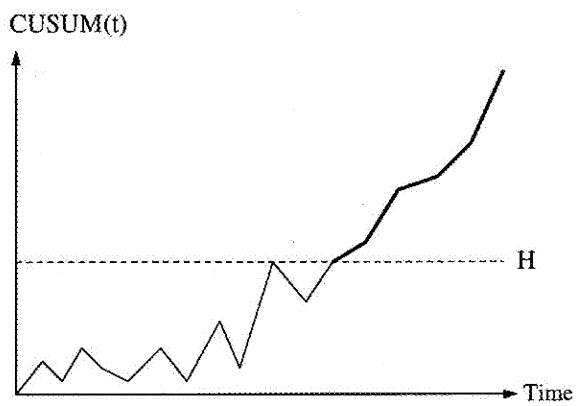

Fig. 4　CUSUM chart

k_H,k_L,HなどのパラメーターはCUSUMをどのようなプロセスに適用させるかで異なる。設定値からの閾値k_H,k_Lは各変数の標準偏差を基準に決定することが多い。

2.6　Postprocessing

他の手法による処理結果を受けて実施する手法や、SPCのような管理値を算出するための手法などが含まれる。

3.　TennesseeEastmanプラント

ツールボックスの適用対象としたTennessee Eastman プラントと、プラント内で発生する異常の種類について説明を行なう。

3.1　プラント概要とフローシート

Tennessee Eastman プラントとはEastman Chemical Companyで運転されていた化学プラントの成分、装置、操作条件などに多少の変更を加え作成された化学プラントモデルシミュレーターである[5]。化学プラントの理解、制御手法のテストなど様々な分野に利用可能であり、FORTRANで書かれたプログラムは無償で配布されている。

Tennessee Eastman プラントのフローシートをFig.5に示す。

Tennessee Eastman プラントは反応器・凝縮器・気液分離器・コンプレッサー・吸収塔の５つの装置、各装置間にある13本のストリーム、そして反応器入口・パージ出口・製品出口の３個所にある組成分析器から構成されている。反応器では4種の反応物(A,C,D,E)によって2種の製品(G,H)を生産している。この4種の反応物はフィードによる供給と、気液分離器からリサイクルによって得られ、運転目標は製品(G,H)を安定生産することである。また成分Bは不活性成分、Fは反応時の副生成物である。

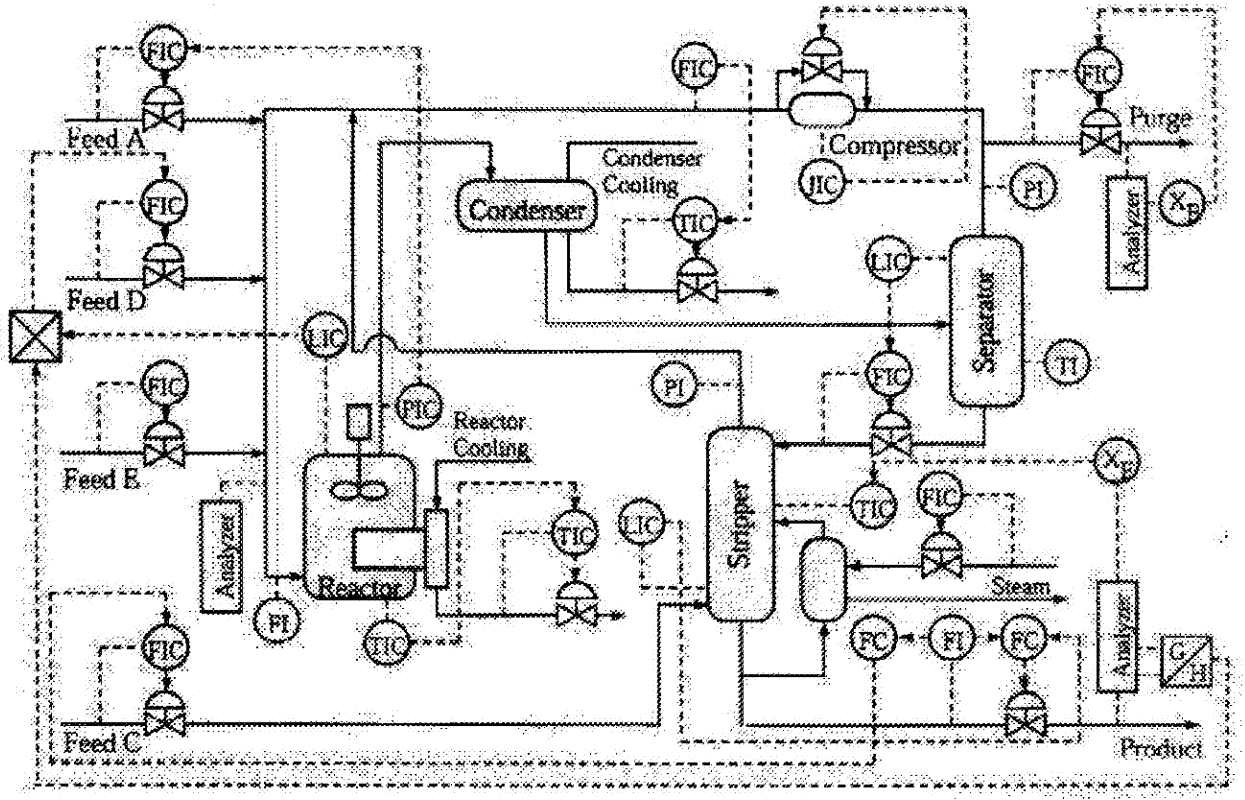

Fig. 5　Tennessee Eastman Process

操作変数として各フィード・装置冷却水のバルブ開度・反応器撹拌速度等、12の変数を操作できる。

計測できる変数としてストリーム・冷却水流量、各装置の温度・圧力・液面レベルなど22の変数がある。組成分析器からは19の組成測定値が得られ、プラント全体で合計53の変数がある。

実際の計算にはPID制御系[6]を作成しプラントデータの作成を行なった。

Table 2　Process disturbances

No.	Disturbances	Type
1	A/C feed ratio	Step
2	B composition	Step
3	D feed temp	Step
4	Reactor cooling water inlet temp	Step
5	Condenset cooling water inlet temp	Step
6	A feed loss	Step
7	C header pressure loss	Step
8	A,B,C feed composition	Random
9	D feed temp	Random
10	C feed temp	Random
11	Reactor cooling water inlet temp	Random
12	Condenser cooling water inlet temp	Random
13	Reaction kinetics	Slow drift
14	Reactor cooling water valve	Sticking
15	Condenser cooling water valve	Sticking
16-20	Unknown	Unknown

3.2　異常データの生成

Tennessee Eastman プラントにはあらかじめ20種の異常（外乱）がプログラムソース中に用意されており、それらを任意に発生させることができる。一覧をTable.2に示す。

Table.2の異常をそれぞれ1種ずつ発生させKDDMツール解析の対象とするためのシミュレーションデータを20個作成した。すべてのデータは8時間の運転を想定して作成した。運転開始から3時間経過時点で異常が発生するとし、異常発生前をクラスNormal、発生後をWarningとした。サンプリング間隔は3分としているため8時間分の運転データ数は161点から成り、異常は61点目から発生する。

データの作成後、前処理としてデータに一次微分を加えたため、この時点でデータ中の属性数は2倍になり106変数となった。その後で移動平均フィルターを適用しノイズを平滑化した、このときの移動平均フィルターの計算に用いた平滑化点数は5点とした。

Table 3　Selected variables

No.	Variable name
1	A feed
2	A and C feed
3	Reactor Pressure
4	Reactor temp
5	Purge rate
6	Product sapareor temp
7	Product saparator pressure
8	Stripper pressure
9	Stripper underflow
10	Stripper temp
11	Stripper steam flow
12	Compressor work
13	Reactor cooling water outlet temp
14	Condenser cooling water outlet temp
15	Reactor B component
16	Purge gas D component
17	Product D component
18	A feed valve
19	Separator pot liquid flow valve
20	Stripper liquid product flow valve
21	Reactor Pressure (differential)
22	Product separator temp (differential)
23	Product separator underflow (differential)
24	Reactor cooling water outlet temp (differential)

3.3　KDDMツールによる解析例

3.3.1　変数選択

Tennessee Eastman プラントに異常が発生したデータに対してKDDMツールによる解析を行なった結果を示す。

まず始めにC4.5を利用してデータマイニングに用いる変数の選択を行なった。作成した異常データに交叉検定を行ない1個のデータから5本の決定木を作成した。この5本の決定木のうち半数以上の決定木に用いられている変数をデータマイニングに採用することにした。これを20個のデータすべてで行なった結果、106変数から24変数(微分値4変数を含む)に絞り込まれた(Table.3)。

そして、選択された変数のみで同様に決定木を作成しテストデータによる決定木の評価を行なった、その結果をTable.4に示す。

変数選択の前後を比較し選択後の誤り率が低下していることから選択前より少ない変数で汎化能力が向上した決定木が作成されているということであり、C4.5での変数選択が有効であると考えられる。

3.3.2　Shewhart chart , Cusumによる検出

Shewhart Chart・Cusumによる異常の検出を試みた。Shewhart chartでは基準値を設置し、そこから上下に基準値の標準偏差の3倍を越えた時に異常と判断する。基準値は正常運転時の平均値とした。

Table 4　Estimation Error for Test Data

DataSetNumber	106Variables	24Variables
1	2.5%	2.5%
2	3.1%	3.1%
3	3.1%	3.1%
4	0.6%	0.6%
5	0.6%	0.6%
6	0.6%	0.6%
7	1.2%	0.6%
8	4.3%	4.3%
9	8.7%	7.4%
10	11.8%	9.3%
11	3.7%	3.7%
12	3.1%	3.1%
13	12.4%	10.6%
14	13.7%	8.1%
15	11.2%	13.1%
16	3.1%	3.7%
17	14.9%	17.4%
18	19.3%	17.4%
19	17.3%	11.2%
20	8.6%	9.9%
Average	7.19%	6.52%

Cusumでは設定値μ_{TR}は正常運転時の平均値を用い、k_H,k_Lはμ_{TR}の標準偏差の2倍を使用、Hは正常データでCusumを計算した最大値とした。この条件下で検出を行なった結果を以下に示す。

No.1～7のStep変化の異常に対しては、C4.5を用いた検出と同様に、Shewhart・Cusumの両方が3時間経過後からすぐに異常を検出した。

No.8～12のRandom変化の異常に対しては、検出にCusumで異常発生から約5点、Shewhartで約10点の遅れが生じた、特にNo.10の異常(Fig.6)に関しては約40点の検出遅れが生じた。

No.13～15ではShewhart・Cusumの両方で約20点の検出遅れが生じた(Fig.7)

No.16～20では異常によって検出に差があり、良

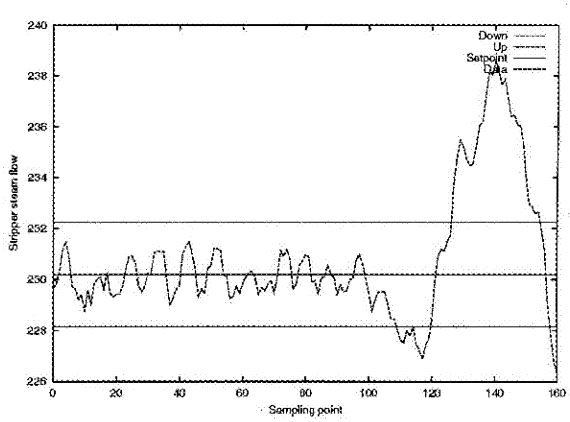

Fig. 6　Shewhart chart (Disturbance No.10)

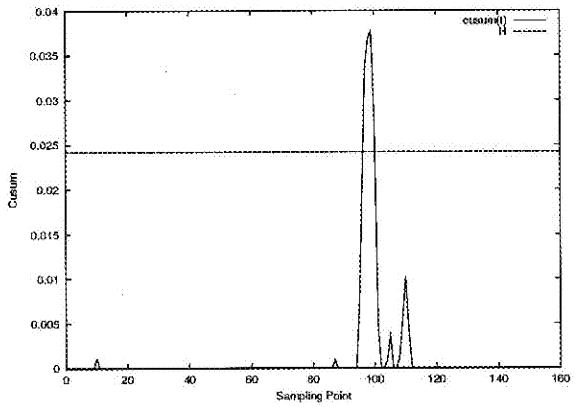

Fig. 7 Cusum chart (Disturbance No.15)

Table 5 Example Rules (No.8)

If	(Perge rate <= 0.338752)
and	(Compressor work > 341.18)
and	(Separator pot liquid flow valve <= 38.6698)
and	(Reactor pressure (differential) > -0.472194)
Then	(Class = Normal)
If	(Purge Rate > 0.338752)
Then	(class = Warning)
If	(Compressor Work <= 341.18)
Then	(class = Warning)
If	(Separator pot liquid flow valve > 38.6698)
Then	(class = Warning)
If	(Reactor pressure (differential) <= -0.472194)
Then	(class = Warning)
Others	(class = Warning)

い結果で約5点の検出遅れ、悪い結果では50点以上の検出遅れが生じた。

3.3.3 If-Thenルールによる検出

選択した24変数で20個の異常データについてそれぞれ決定木を作成し、その後C4.5rulesを用いてIf-Thenルールとした。その一部をTable.5に示す。異常データにこのIf-Thenルールを適用して異常の検出を行なってみた。その結果の一部を以下に示す。この例ではt=61から異常状態である。図でy軸の0は正常(Normal)と、1は異常(Warning) と判断をしたことを表している。

20種の異常のうち、No.1～7のStep変化の異常に関しては、すべてが異常が発生する3時間経過後からIf-Thenルールによって異常を正しく検出できている(Fig.8)。

Random変化を行なうNo.8～12ではNo.10以外の異常では正しい検出を行なったが、No.10(Fig.9)では3時間経過する前に異常と判断してしまうミスが発生していた。

No.13～15では、異常発生前に異常と判断するミスに加えて異常発生後に正常と判断してしまうミスが発生(Fig.10)しており、検出の難しい異常であると言える。

No.16～20では、全体的に良好な検出が行なえた。

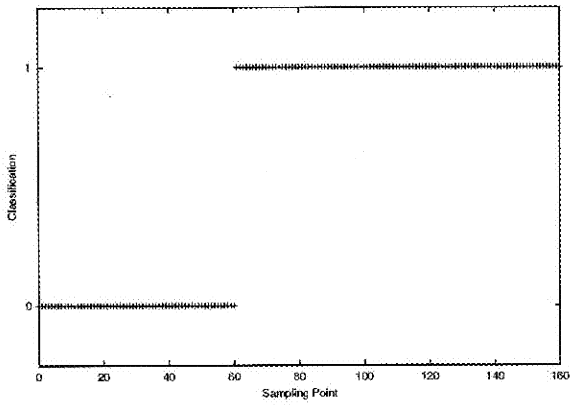

Fig. 8 Fault Detection by Rules (Disturbance No.1)

4. 結言

KDDMツールボックスを試作し、Tennessee Eastman プラントの異常データに対して適用した、C4.5から得たIf-Thenルールによる検出とShewhart Chart・Cusumによる検出を比較した結果、特にRandom変化をする異常に関してはC4.5による検出がChartによる検出より有効であったと言える。

Step変化の異常の検出は両方とも異常発生後すぐ検出することができたが、No.13～15の異常では両方の検出に欠点があることから、異常の発生の種類が検出性能に影響を与えていることが分かった。

またIf-Thenルールでの検出ではIf-Thenルールに適用した結果のみをチェックすれば良いのに対し、チャートの場合は観測している変数すべてを

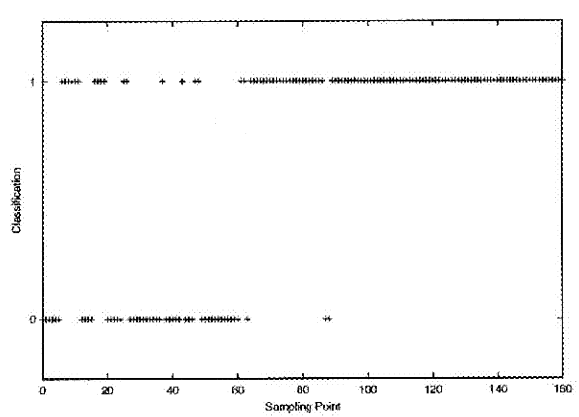

Fig. 9 Fault Detection by Rules (Disturbance No.10)

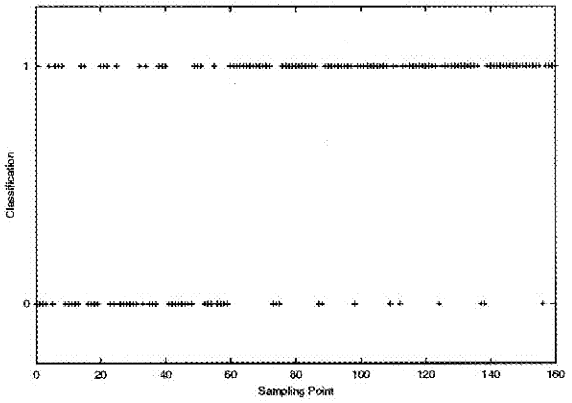

Fig. 10 Failt Detection by Rules (Disturbance No.15)

チェックしなければならず利便性も含めC4.5が有効であると言える。

今回は複数のKDD手法を統合的に適用するまでには至らなかったが、今後はツールボックスに新たな機能を加えるとともに、適用例を通じてその有効性を検証して行きたい。

参考文献

1) Yoshiyuki Yamashita : Supervised learning for the analysis of process operational data , Comput . Chem . Eng **24** , 471/474 (2000)

2) Yoshiyuki Yamashita : Data Based Approach for Intelligent Process Monitoring , PSE Asia 2000 - International Symposium on Design,Operation and Control of Next Generation Chemical Plants, Kyoto, December (2000)

3) The MaTX Home Page, http://www.matx.org/

4) J,R,Quinlan,古川康一監訳: AIによるデータ解析, トッパン(1995)

5) J.J.Downs and E.F.Vogel:A Plant-Wide Industrial Process Control Problem,Comput.Chem.Eng, **3**, 245/255 (1993)

6) T,J,McAvoy and N.Ye: Base Control For Tennessee Eastman Problem,Comput.Chem.Eng, **5**, 383/413 (1994)

7) 日本学術振興会,プロセスシステム工学第143委員会:知的モニタリング -時系列データからの情報抽出-