ソフトセンサーのためのデータベース管理指標の開発

金子 弘昌^a、船津 公人^{a*}

^a東京大学大学院工学系研究科 (〒113-8656 東京都文京区本郷7-3-1)

(Received March 16 2013; Accepted May 1, 2013)

産業プラントにおいては測定困難なプロセス変数をリアルタイムに推定する手法としてソフトセンサーが広く使用されている。触媒の劣化や原料組成の変化等によりプラントの運転状態は変化するため、ソフトセンサーモデルはそのプロセス特性の変化に追随しなければならない。そこで新しく測定された運転データを活用したモデルの再構築が行われている。しかしばらつきが小さいデータを用いてモデルが再構築された場合、そのモデルはプロセスの急激な変化に対応できない。モデルの予測性能はデータベースに依存するといえる。そこで本研究では、データベースを管理するための指標database monitoring index (DMI)およびDMIを用いたデータベース管理手法を提案する。DMIは2つのデータの類似度に基づく指標であり、目的変数の差の絶対値と説明変数の類似度との比で定義される。2つのデータが類似しているほどDMIの値は小さくなる。新しいデータが測定された際、そのデータとデータベース内の全データとのDMI値を計算し、それらの最小値が大きいデータのみデータベースに追加する。これによりデータベースのデータ数を抑えながらデータベースの情報量を大きくすることができる。本手法を用いてシミュレーションデータ解析を行った結果、DMIにより適切なデータベース管理が可能であり、ソフトセンサーモデルの予測性能が向上することを確認した。

キーワード: ケモメトリックス、プロセス管理、ソフトセンサー、データベース、モニタリング

1. 緒言

産業プラントを安全・安定的・効率的に運転するため、温度・圧力・流量・濃度などのプロセス変

数の値を測定して適切に管理しなければならない。温度・圧力・流量などの測定は容易であり、リアルタイムに測定値を得ることができるが、濃度のように頻繁な測定が困難であったり、測定に時間がかかったりするプロセス変数も存在する。そのようなプロセス変数のリアルタイムな制御は難しく、制御に

* funatsu@chemsys.t.u-tokyo.ac.jp

時間遅れが生じてしまう。そこで測定が困難なプロセス変数の値を推定する手法としてソフトセンサー [1,2] が利用されている。ソフトセンサーとは、これまでに測定された運転データを用いて温度・圧力・流量などの容易に測定可能な変数 $\mathbf{X}$ と濃度などの測定困難な変数 $\mathbf{y}$ の間で構築された数値モデルである。物理的・化学的な背景や知識を方程式化してモデルを構築することもあるが、統計手法により構築された回帰モデルを用いることの方が多い。 $\mathbf{X}$ の値をモデルに入力することで、 $\mathbf{y}$ の値をリアルタイムに推定することができる。推定値を実測値の代替として使用することで、迅速かつ効率的な制御が達成され大きなコストの削減となる。

しかしソフトセンサーには問題点も存在する。重要な問題点の一つにモデルの劣化 [3] が挙げられる。モデルの劣化とは、原料組成の変化・配管や反応器等への汚れ付着・触媒の劣化などの外乱によってプラントの運転状態が変化し、モデルの予測精度が低下してしまう現象である。

モデルの劣化を低減するため、最新の運転データを活用してプラントの状態変化にモデルを追従させる試みがなされている。このように最新のプラント状態に適応しながら $\mathbf{y}$ の値を予測するモデルは適応型モデルと呼ばれ、多数の研究例が報告されている [4]。この適応型モデルは大きく三つに分類される。時間差分モデル (time difference, TD) モデル [5]・moving window (MW) モデル [6]・just-in-time (JIT) モデル [7]である。TDモデルとは、各プロセス変数の時間差分の間で構築されたモデルである。複数の差分間隔による予測値を用いることで予測精度の向上および予測誤差の推定も可能である [8]。MWモデルは予測したいデータに時間的に近いデータセットのみを用いて構築されたモデルである。JITモデルは予測データに近いデータを用いたり、予測データに近いデータほど重みを大きくしたりして構築されたモデルである。類似度の指標としてデータ間のユークリッド距離や相関係数などが用いられる。MWモデルおよびJITモデルにおけるモデルの再構築のため、新しく測定されたデータはデータベースに追加され、古いと判断されたデータはデータベースから削除される。

すべてのプラント状態において万能に使用可能な適応型モデルは存在せず、各状態によって適応型モデルの精度の優劣は変化する [9]。金子らはモデル劣化の種類を分類し、その分類結果に基づいて適応型モデルの特徴を整理した [10]。 $\mathbf{X}$ または $\mathbf{y}$ の値がシフトする劣化が起きた際は、劣化の起きる速さに関わらずTDモデルで対応可能である。しかしTDモデルは $\mathbf{X}$ と $\mathbf{y}$ の傾きの変化に対応できない。一方で $\mathbf{X}$ と $\mathbf{y}$ の傾き変化が起きた場合、 $\mathbf{X}$ の値がシフトしてい

ればJITモデルを使用すべきであるが、シフトしていなければJITモデルでもその劣化には対応不可能である。ただMWモデルとして非線型回帰分析手法と時刻を表す変数を組み合わせることで、 $\mathbf{X}$ の値シフトがない場合でも傾きの急激な変化に対応できることが確認されている [11]。

このようにJITモデルやMWモデルを上手く使用することで $\mathbf{X}$ と $\mathbf{y}$ の傾き変化に対応できるが、それらのモデルには問題点も存在する。その中の一つが、モデル構築用データにあまりばらつきがない場合に、その後起こるプロセスの急激な変化にモデルが対応できないことである [5]。例えば変動の少ないデータでMWモデルが更新され続けると、その後に大きなプロセス変動が起きた際に予測精度が低下してしまう。ただ一方で、プロセスやセンサーが時間的にゆっくりとドリフトする場合も多く、モデルを更新しないとドリフトの影響を受けてモデルが劣化してしまう。

そこで本研究では、幅広いデータ範囲において予測精度の高い適応型モデルを構築するために、適切なデータベース管理を行うことを目的とした。データベース内のデータ数が多過ぎるとMWモデルやJITモデルの構築に時間がかかってしまうため、新しく測定されたデータをデータベースに追加するかどうか検討することが必要となる。情報量が多いデータは追加すべきであり、情報量が少ないデータは追加しなくて良いといえる。

本論文では、データベースを管理するための指標 database monitoring index (DMI) およびDMIを用いたデータベース管理手法を提案する。DMIは2つのデータの類似度に基づく指標であり、 $\mathbf{y}$ の差の絶対値と $\mathbf{X}$ の類似度との比で定義される。2つのデータが類似している場合にDMIの値は小さくなり、異なる場合にDMI値は大きくなる。新しいデータに対してデータベース内の全データとのDMI値を計算し、その最小値が大きいデータのみをデータベースに追加することで、データベースのデータ数を抑えながら情報量を増やすことができる。DMIを用いることで、 $\mathbf{X}$ および $\mathbf{y}$ の値シフトだけでなく、 $\mathbf{X}$ と $\mathbf{y}$ の傾きの変化を考慮したデータベース管理が可能となる。

本手法の有効性を確認するため、 $\mathbf{X}$ と $\mathbf{y}$ が非線型関係を持っており、ある一定時間の変動が小さい数値シミュレーションデータを用いた解析を行う。 $\mathbf{y}$ がドリフトする場合としない場合で検討する。DMIを用いたデータベース管理を行うことで、 $\mathbf{X}$ および $\mathbf{y}$ の変動が小さい状態が続いた後に急激なプロセス特性の変化があった場合でも、適応型モデルがその変化に追従可能であることを示す。

2. 手法

DMI

本論文で提案するデータベース管理のための指標DMIは2つのデータ($\mathbf{x}_i, y_i$), ($\mathbf{x}_j, y_j$)の間で以下の式により定義される。

$$\text{DMI} = \frac{|y_i - y_j|^a}{\text{sim}(\mathbf{x}_i, \mathbf{x}_j)} \quad (1)$$

ここで $\text{sim}(\mathbf{x}_i, \mathbf{x}_j)$ は $\mathbf{x}_i$ と $\mathbf{x}_j$ の類似度、 a はある定数である。類似度として、例えばユークリッド距離やマハラノビス距離の逆数・相関・ガウシアンカーネルなどのカーネル関数が挙げられる。本研究において類似度として使用したガウシアンカーネルを以下に示す。

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \quad (2)$$

ここで γ はカーネル中のパラメータを表す。よって今回用いたDMIは以下の式で表される。

$$\text{DMI} = \frac{|y_i - y_j|^a}{\exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)} \quad (3)$$

DMIの概念図をFigure 1に示す。DMIの値は、 $\mathbf{X}$ および $\mathbf{y}$ が類似していると小さくなり、相違していると

Figure 1. The basic concept of DMI.

大きくなる。

a の値を変化させることで $\mathbf{X}$ に対する $\mathbf{y}$ の重みを変更可能である。 $|y_i - y_j|$ と $|y_i - y_j|^a$ の関係をFigure 2に示す。 a の値が1より小さいほど、例えばドリフトのような y の微小変化に対して $|y_i - y_j|^a$ が大きくなりDMIの値も大きくなる。ただし y のノイズの影響をDMIが受けやすいため注意が必要である。一方 a の値が1より大きいと y の変化が小さい時にDMI値が小さくなる。 y のドリフト時にDMI値が小さくなってしまいう一方で、DMIは y のノイズの影響を受けにくくなる。 a の値を変化させることで対象プロセスの状態およびデータベース管理の目的に応じたDMIを検討できる。

DMIによるデータベース管理

DMIを用いたデータベース管理の流れをFigure 3に示す。新規データが得られた際、データベースに含まれる全データと新規データの間でDMIを計算する。それらの最小値がある閾値 P_{DMI} を超えた場合はそのデータをデータベースに追加するが、 P_{DMI} 以下の場合はデータベースに追加しない。

DMIを用いることで、MWモデルの構築に用いるデータの管理だけでなく、JITモデルにおけるデータベース管理を行うことが可能である。距離や相関などの指標でデータを選択するJITモデルでは多くのデータからモデル構築用データを選択しなければならず、データに重み付けを行うJITモデルでは多くのデータに対して重みを付けなければならない。データベースを適切に管理することで効率的に精度の高いJITモデルを構築できるようになる。

(1)式のような $\mathbf{X}$ と $\mathbf{y}$ の類似度の比を用いた指標として、構造活性相関の分野ではstructure-activity

Figure 2. The relationships between $|y_i - y_j|$ and $|y_i - y_j|^a$.

Figure 3. The flow of monitoring database with DMI.

landscape index (SALI) [12]が有名である。2つの分子 p および q に対してSALIは以下のように表される。

$$\text{SALI} = \frac{|A_p - A_q|}{1 - \text{sim}(p, q)} \quad (4)$$

ここで A_p と A_q はそれぞれ p と q の活性値、 $\text{sim}(p, q)$ は p と q との化学構造の類似度である。 p と q の活性値が異なり化学構造の類似度が大きいとSALIの値が大きくなる。SALIを活用することで、activity cliffと呼ばれる化学構造は類似しているが活性が大きく異なる化合物の組の検出など多くの成果が挙げられている [13]。

3. 結果と考察

データ

DMIの有効性を確認するため、 $\mathbf{X}$ と $\mathbf{y}$ の関係が非線型の場合を想定した数値シミュレーションデータを用いた解析を行った。 $\mathbf{X}$ は2変数 x_1, x_2 であり、 x_1, x_2, y の関係は以下の式で表される。

$$y = \sin(x_1)\cos(x_2) + 0.1x_1 \quad (5)$$

(5)式は文献 [14] のテスト問題として掲載されている式に $0.1x_1$ を追加した式である。 x_1, x_2, y の関係を

Figure 4に示す。カラーバーは y の値に対応している。

$\mathbf{X}$ のデータをランダムウォークにより発生させ、 x_1, x_2 ともに ± 3 の間にレンジスケリングした。 x_1, x_2, y には平均0、標準偏差0.01の正規乱数を追加した。 x_1, x_2, y の時間プロットをFigure 5に示す。時刻785から200データはノイズのみの変動であり、その様子はFigure 5からも確認できる。最初の100データをトレーニングデータ、その後の1,200データをテストデータとした。

今回は時間的に一定に y の値が低下するドリフトが存在する場合の検討も行う。ドリフトのみの標準偏差が y の全データの標準偏差の0.01倍になるよう重み付けして y に追加した。

y にドリフトがない場合

カーネル関数をガウシアンカーネルとした support vector regression (SVR) 法 [15] を用いて、5-fold cross validationにより γ の値を最適化した。この γ の値 2^{-5} を用いて(3)式によりDMIを計算し、Figure 3の流れでデータベースの更新を行った。データベースに含まれるデータ数の上限を50と仮定し、古いデータは自動的に削除した。 a を0.5、1、2として P_{DMI} を0から0.0001ずつ1まで変化させ、それぞれデータベースを更新した割合を計算した。結果をFigure 6に示す。 P_{DMI} が大きくなるにつれてデータベースを更新する割合が小さくなること、および a を変えることで P_{DMI} とデータベースを更新する割合の関係が変化することが確認できる。

続いて適応型モデルによる y の値の予測を行った。今回は適応型モデルとしてonline support vector regression (OSVR) [11,16] モデルを使用する。OSVRの詳細についてはAppendix Aを参照されたい。 y の測定にかかる時間を考慮するため、ある時刻の $\mathbf{X}$ の値

Figure 4. The relationship among x_1, x_2 and y . The colorbars represent y -values.

Figure 5. The time plots of x_1 , x_2 and y .

は瞬時に得られるが y の値は時刻が5進んだ後に得られるとした。 y の値が得られた後に、そのデータをデータベースに追加するかどうかDMIと P_{DMI} により検討する (Figure 3)。

P_{DMI} を変化させてOSVRモデルにより予測を行った結果をTable 1に示す。 r_p^2 はテストデータの決定係数 r^2 、 $RMSE_p$ はテストデータのroot-mean-square error (RMSE)である。 r_p^2 が1に近いほど、 $RMSE_p$ が0に近いほどモデルの予測精度は高い。Table 1より P_{DMI} が0の毎時間データがデータベースに追加されモデルが更新される場合と比較して、少ない更新頻度で精度の高い予測が可能であることを確認した。 r_p^2 値が最大となり $RMSE_p$ 値が最小となった P_{DMI} が0.07の結果から、データベースに追加するデータを適切に選択することで12.0%の更新のみで精度良く予測できたことになる。 $\mathbf{X}$ と $\mathbf{y}$ の非線型性を表現するために必要なデータを提案したDMIで適切に選択可能であった。

Figure 7に P_{DMI} を0, 0.001, 0.07とした際の、時刻950から1050までの y の時間プロットを示す。Figure 7aより P_{DMI} が0の毎時刻モデルが更新された場合に時刻990付近の急激な変動時の予測誤差が大きくなったことが分かる。変動のないデータのみでモデルが更新され、その状態の予測に特化してしまい、後の急激な変動にモデルが追従できなかった。一方

P_{DMI} を0.001としてモデルを更新するデータを選択して変動のないデータでのモデル更新を避けることで、時刻990付近の急激な変動に適切に追従可能であった (Figure 7b)。さらにFigure 7cより更新の割合が12.0%である P_{DMI} が0.07の際も精度良く予測できていることが分かる。提案する指標を用いてデータベースに追加するデータを適切に選択することで、幅広い y の範囲を高い精度で予測可能であることを確認した。

Figure 6. The relationship between P_{DMI} and rate of update when no drift of y existed.

ドリフトがある場合

y のみにドリフトの影響を与えてこれまでと同様の検討を行った。 a を変化させた際の P_{DMI} とデータベースを更新する割合との関係をFigure 8に示す。ドリフトがない場合と同様に、 P_{DMI} が大きくなるととも

に更新の割合は減少し、また a の変化に伴い P_{DMI} と更新の割合との関係も変わることを確認した。

Table 2にOSVRモデルによる予測結果を示す。ドリフトの影響がある場合も、適切に P_{DMI} を選択することで毎回モデルを更新する P_{DMI} が0の場合と比較して r_p^2 の値が大きく $RMSEP$ の値が小さくなることを

(a) $P_{\text{DMI}} = 0$

(b) $P_{\text{DMI}} = 0.001$

(c) $P_{\text{DMI}} = 0.07$

Figure 7. The time plots of simulated and predicted y from 950 to 1050 when no drift of y existed.

Figure 8. The relationship between P_{DMI} and rate of update when drift of y existed.

確認した。 r_p^2 値が最大で $RMSEP$ 値が最小となったのは P_{DMI} が0.002の場合であり、更新する割合は79.9%であった。ドリフトによる $\mathbf{X}$ と $\mathbf{y}$ の関係性の変化にモデルが追随する必要があるため、ドリフトの影響が

ない場合と比較して更新する回数が増えたと考えられる。

P_{DMI} を0、0.002、0.07とした際の、時刻950から1050までの $\mathbf{y}$ の時間プロットをFigure 9に示す。データが得られるごとに毎回モデルを更新した場合は、変動の小さいデータに適合するようモデルが構築されてしまい、その後の急激な変動に追随できずに予測誤差が大きくなってしまった (Figure 9a)。一方でFigure 9bおよびFigure 9cより、DMIを指標にしてデータベースに追加してモデル更新するデータを選択することで、変動の小さい時間だけでなくその後に急激な変動が起きた際も精度良く予測できたことが分かる。ドリフトの影響がある場合でも提案した指標により適切なデータベース管理が可能であった。

今回は P_{DMI} を変化させて検討を行ったが、実際にDMIによるデータベース管理を実施する際は事前に適切な P_{DMI} を決定しなければならない。

Table 1. The prediction results for each P_{DMI} when no drift of y existed.

P_{DMI}	Rate of update	r_p^2	$RMSEP$
0	1.0000	0.9962	0.0378
0.0001	0.9775	0.9964	0.0367
0.0002	0.9517	0.9961	0.0381
0.0005	0.8975	0.9972	0.0323
0.0007	0.8675	0.9978	0.0288
0.001	0.8317	0.9976	0.0300
0.002	0.7667	0.9975	0.0307
0.005	0.6500	0.9971	0.0330
0.007	0.5900	0.9976	0.0301
0.01	0.5200	0.9970	0.0333
0.02	0.3733	0.9975	0.0305
0.05	0.1675	0.9968	0.0343
0.07	0.1200	0.9980	0.0270
0.1	0.0825	0.9930	0.0508
0.2	0.0358	0.9902	0.0603
0.5	0.0108	0.8789	0.2121
0.7	0.0058	0.7869	0.2814
5	0.0000	-0.1407	0.6510

Table 2. The prediction results for each P_{DMI} when drift of y existed.

P_{DMI}	Rate of update	r_p^2	$RMSEP$
0	1.0000	0.9958	0.0416
0.0001	0.9850	0.9957	0.0424
0.0002	0.9642	0.9956	0.0427
0.0005	0.9233	0.9955	0.0432
0.0007	0.9025	0.9959	0.0413
0.001	0.8675	0.9966	0.0374
0.002	0.7992	0.9971	0.0344
0.005	0.6817	0.9971	0.0344
0.007	0.6208	0.9969	0.0359
0.01	0.5633	0.9964	0.0385
0.02	0.4233	0.9965	0.0382
0.05	0.2208	0.9958	0.0416
0.07	0.1725	0.9959	0.0413
0.1	0.1258	0.9954	0.0434
0.2	0.0675	0.9898	0.0649
0.5	0.0267	0.9708	0.1099
0.7	0.0183	0.9717	0.1083
5	0.0000	0.8773	0.2253

4. 結言

本研究では適応型ソフトセンサーを運用する際に適切にデータベースを管理するため、DMIという

指標およびDMIを用いたデータベース管理手法を提案した。新規データについて、そのDMIの値でデータベースに追加するかしないかを判断する。必要なデータのみデータベースに追加することで、データ数を抑えながらデータベース内の情報量を増加さ

Figure 9. The time plots of simulated and predicted y from 950 to 1050 when drift of y existed.

せることが可能となる。数値シミュレーションデータ解析を通して、DMIを指標にして適切にデータベースの管理が可能であり、適応型モデルの予測性能が向上することを確認した。

今回は適応型モデルの中でMWモデルのみ検討を行ったが、JITモデルを使用する際も適切なデータベース管理が必要である。JITモデルを構築する際、予測データに近いデータをデータベースから選択する方法や予測データに近いデータほどデータベース内のデータの重みを大きくする方法がある [17]。新しい運転データをデータベースに蓄積しないと最新のプラント状態に追従できないが、すべてをデータベースに保存するとデータベースのサイズが大きくなりJITモデルの構築に多くの計算時間がかかってしまう。提案手法を用いてFigure 3の流れによりデータベースに蓄積するデータを取捨選択することで適切にJITモデルを運用可能といえる。

実際はDMIの α および閾値 P_{DMI} を事前に設定しなければならない。これは今後の課題の一つである。今回は最初のモデル構築用データはすべてデータベースに蓄積されているとしたが、トレーニングデータ間のDMIの値を計算して確認することで情報量の少ないデータをデータベースから削除できると考えられる。これにより情報量を保ちながらコンパクトなデータベースを構築可能である。また本論文では新しい測定データに異常値はないと仮定したが、実際は異常値がデータベースに混入すると適応型モデルの予測性能は低下してしまう。適切な異常値検出 [3]もデータベース管理には必須項目である。

本手法によりデータベースを適切に管理してソフトセンサーの予測精度が向上することで、迅速なプロセス制御が達成され安定かつ効率的なプラント管理が実現されると期待できる。

謝辞

本研究の一部は、金子弘昌が科研費の若手研究(B)(No. 24760629)および財団法人みずほ学術振興財団の助成を受けたものである。ここに深く謝意を示す。

参考文献

- [1] M. Kano, Y. Nakagawa, *Comput Chem Eng.*, **32**, 12–24, 2008.
- [2] P. Kadlec, B. Gabrys, S. Strandt, *Comput Chem Eng.*, **33**, 795–814, 2009.
- [3] H. Kaneko, M. Arakawa, K. Funatsu, *AIChE J.*, **55**, 87–98, 2009.
- [4] P. Kadlec, R. Grbic, B. Gabrys, *Comput. Chem. Eng.*, **35**, 1–24, 2011.
- [5] H. Kaneko, K. Funatsu, *Chemom. Intell. Lab. Syst.*, **107**, 312–317, 2011.
- [6] P. Kadlec, B. Gabrys, *AIChE J.*, **57**, 1288–1301, 2010.
- [7] C. Cheng, M. S. Chiu, *Chem. Eng. Sci.*, **59**, 2801–2810, 2004.
- [8] H. Kaneko, K. Funatsu, *Chemom. Intell. Lab. Syst.*, **109**, 197–206, 2011.
- [9] T. Okada, H. Kaneko, K. Funatsu, *Songklanakarin J. Sci. Technol.*, **34**, 217–222, 2012.
- [10] H. Kaneko, K. Funatsu, *AIChE J.*, in press.
- [11] 金子 弘昌, 船津 公人, 第13回計測自動制御学会制御部門大会, 7C2-2, 2013.
- [12] R. Guha, J. H. V. Drie, *J. Chem. Inf. Model*, **48**, 646–658, 2008.
- [13] D. Stumpfe, J. Bajorath, *RSC Adv.*, **2**, 369–378, 2012.
- [14] G. Li, V. Aute, S. Azarm, *Struct. Multidiscip. O.*, **40**, 137–155, 2010.
- [15] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
- [16] J. Ma, J. Theliler, S. Perkins, *Neural Comput.*, **15**, 2683–2703, 2003.
- [17] M. Kano and K. Fujiwara, *J. Chem. Eng. Japan*, **46**, 1–17, 2013.

Appendix A: OSVR

OSVR 法とは、SVR モデルが満たすべき Karush-Kuhn Tucker (KKT) 条件について、データが追加および削除された際にも成立するよう効率的にSVRモデルを更新する手法である。SVR法においては、下記の値を最小化するように学習が行われる。

$$\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i |y_i - f(\mathbf{x}_i)|_{\varepsilon} \quad (\text{A.1})$$

ただし、

$$|y_i - f(\mathbf{x}_i)|_{\varepsilon} = \max(0, |y_i - f(\mathbf{x}_i)| - \varepsilon) \quad (\text{A.2})$$

である。ここで y_i と $\mathbf{x}_i$ は i 番目のモデル構築用データ、 f はSVRモデル、 C は2つの項の間で重み付けを調整するための係数、 $\mathbf{w}$ はウェイトベクトル、 ε は誤差の閾値である。あるデータ $\mathbf{x}$ の予測値は以下の式で表

される。

$$f(\mathbf{x}) = \sum_{i=1}^N (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \quad (\text{A.3})$$

ここで N はモデル構築用データ数、 b は定数項、 K はカーネル関数である。本研究ではカーネル関数として以下に示すガウシアンカーネルを使用した。

$$K(\mathbf{x}_i, \mathbf{x}) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2\right) \quad (\text{A.4})$$

γ はカーネル中のパラメータである。(A.1)(A.2)式より、(A.3)式中の α_i 、 α_i^* は下記の値を最小化することにより得られる。

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N K_{ij} (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) \\ & - \sum_{i=1}^N y_i (\alpha_i - \alpha_i^*) + \varepsilon \sum_{i=1}^N (\alpha_i + \alpha_i^*) \end{aligned} \quad (\text{A.5})$$

ただし、

$$0 \leq \alpha_i, \alpha_i^* \leq C \quad i=1, 2, \dots, N \quad (\text{A.6})$$

$$\sum_{i=1}^N (\alpha_i - \alpha_i^*) = 0 \quad (\text{A.7})$$

$$K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) \quad (\text{A.8})$$

である。ここで、

$$\theta_i = \alpha_i - \alpha_i^* \quad (\text{A.9})$$

とすると、(A.3)(A.4)(A.8)式よりあるデータ $\mathbf{x}_i$ の予測値は以下のように表現される。

$$f(\mathbf{x}_i) = \sum_{j=1}^N K_{ij} \theta_j + b \quad (\text{A.10})$$

ただし(A.7)式より、

$$\sum_{i=1}^N \theta_i = 0 \quad (\text{A.11})$$

である。誤差関数 h を、

$$\begin{aligned} h(\mathbf{x}_i) & \equiv f(\mathbf{x}_i) - y_i \\ & = \sum_{j=1}^N K_{ij} \theta_j + b - y_i \end{aligned} \quad (\text{A.12})$$

と定義すると、KKT条件は下記の式にまとめることができる。

$$h(\mathbf{x}_i) > \varepsilon, \quad \theta_i = -C \quad (\text{A.13})$$

$$h(\mathbf{x}_i) = \varepsilon, \quad -C < \theta_i < 0 \quad (\text{A.14})$$

$$|h(\mathbf{x}_i)| < \varepsilon, \quad \theta_i = 0 \quad (\text{A.15})$$

$$h(\mathbf{x}_i) = -\varepsilon, \quad 0 < \theta_i < C \quad (\text{A.16})$$

$$h(\mathbf{x}_i) < -\varepsilon, \quad \theta_i = C \quad (\text{A.17})$$

各モデル構築用データは(A.13)-(A.17)式のどれかに当てはまることになる。(A.13)(A.17)式を満たすデータをerror support vector (E)、(A.14)(A.16)式を満たすデータをmargin support vector (S)、(A.15)式を満たすデータをremaining vector (R)と呼ぶ。

新しいデータ $(\mathbf{x}_c, y_c)$ が追加される場合を考える。新しいデータがRであれば、SVRモデル θ_i 、 b を更新しなくて良い。EまたはSであれば、新しいデータの θ_c の初期値を0として、 θ_c 、 θ_i 、 b をKKT条件を満たすように少しずつ変化させる。それらの変化により、各データは他の領域に移動する可能性がある。ただ、そのような移動がないと仮定した場合、 $h(\mathbf{x}_i)$ 、 θ_c 、 θ_i 、 b の変化量 $\Delta h(\mathbf{x}_i)$ 、 $\Delta \theta_c$ 、 $\Delta \theta_i$ 、 Δb は(A.11)(A.12)式より以下のように表される。

$$\begin{aligned} \Delta h(\mathbf{x}_i) & = K_{ic} \Delta \theta_c + \sum_{j=1}^N K_{ij} \Delta \theta_j \\ & \quad + \Delta b \end{aligned} \quad (\text{A.18})$$

$$\Delta \theta_c + \sum_{j=1}^N \Delta \theta_j = 0 \quad (\text{A.19})$$

(A.13)(A.15)(A.17)式よりEおよびRのデータでは θ_i が変化しないため(A.18)式は

$$\Delta h(\mathbf{x}_i) = K_{ic}\Delta\theta_c + \sum_{j \in S} K_{ij}\Delta\theta_j + \Delta b \quad (\text{A.20})$$

と変形できる。

(A.14)(A.16)式よりSのデータでは $h(\mathbf{x}_i)$ が変化しないため

$$\sum_{j \in S} K_{ij}\Delta\theta_j + \Delta b = -K_{ic}\Delta\theta_c \quad \forall i \in S \quad (\text{A.21})$$

$$\sum_{j \in S} \Delta\theta_j = -\Delta\theta_c \quad (\text{A.22})$$

となる。(A.21)(A.22)式より、 $\Delta\theta_c$ 、 $\Delta\theta_i$ 、 Δb は以下のように表現できる。

$$\Delta b = \delta\Delta\theta_c \quad (\text{A.23})$$

$$\Delta\theta_i = \delta_i\Delta\theta_c \quad \forall i \in S \quad (\text{A.24})$$

ただし、

$$\begin{bmatrix} \delta \\ \delta_{S_1} \\ \vdots \\ \delta_{S_M} \end{bmatrix} = - \begin{bmatrix} 0 & 1 & \cdots & 1 \\ 1 & K_{S_1 S_1} & \cdots & K_{S_1 S_M} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & K_{S_M S_1} & \cdots & K_{S_M S_M} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ K_{S_1 c} \\ \vdots \\ K_{S_M c} \end{bmatrix} \quad (\text{A.25})$$

$$\delta_i = 0 \quad \forall i \notin S \quad (\text{A.26})$$

である。ここで M は S のデータ数を表す。(A.20)(A.23)(A.24)式より、EおよびRにおける $\Delta h(\mathbf{x}_i)$ は下記のように変形される。

$$\begin{aligned} \Delta h(\mathbf{x}_i) &= K_{ic}\Delta\theta_c + \sum_{j \in S} K_{ij}\Delta\theta_j + \Delta b \\ &= K_{ic}\Delta\theta_c + \sum_{j \in S} K_{ij}\delta_j\Delta\theta_c + \delta\Delta\theta_c \\ &= \left(K_{ic} + \sum_{j \in S} K_{ij}\delta_j + \delta \right) \Delta\theta_c \\ &= \gamma\Delta\theta_c \end{aligned} \quad (\text{A.27})$$

ただし、

$$\gamma = K_{ic} + \sum_{j=1}^N K_{ij}\delta_j + \delta \quad (\text{A.28})$$

である。

(A.24)(A.27)式より各データにおける $\Delta\theta_c$ は、

$$\Delta\theta_c = \delta_i^{-1}\Delta\theta_i \quad \forall i \in S \quad (\text{A.29})$$

$$\Delta\theta_c = \gamma^{-1}\Delta h(\mathbf{x}_i) \quad \forall i \notin S \quad (\text{A.30})$$

となる。(A.29)(A.30)式によりデータごとに、現状の領域から別の領域に移動するための $\Delta\theta_i$ の絶対値の最小値を計算する。つまりEのデータであればSに、SのデータであればEまたはRに、RのデータであればSに移動しうる。各データで計算された $\Delta\theta_i$ の中で絶対値が最小となるデータを実際に移動する。この $\Delta\theta_i$ の絶対値の計算および $\Delta\theta_i$ の中で絶対値が最小となるデータの移動を、全データが(A.13)-(A.17)式のKKT条件を満たすまで繰り返し行う。モデル構築用データからデータを除去する場合も同様の繰り返し計算を行い、KKT条件を満たした時点で終了する。

Development of a New Index to Monitor Database for Soft Sensors

Hiromasa Kaneko^a, Kimito Funatsu^{a*}

^aThe University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan

In industrial plants, soft sensors have been widely used to estimate difficult-to-measure process variables online. The predictive accuracy of soft sensors decreases due to changes in the state of chemical plants, and soft sensor models must adapt to the process changes by using new measured data. However when a model is reconstructed with data that have low variation, the model cannot predict abrupt changes of process characteristics. The predictive performance of adaptive models depends on databases. We therefore propose an index to monitor database, i.e. database monitoring index (DMI), and a database monitoring method using the DMI. The DMI is based on similarity between two data and is defined as a rate of absolute difference of an objective variable and similarity of explanatory variables. The more similar two data are, the smaller value the DMI has. When new data is obtained, DMI values are calculated between new data and all data in a database. If the minimum value of the DMI values is large, the new data is added to the database. By using the DMI and selecting new data, the amount of information of a database can enlarge while curbing the rise in the number of data in the database. Through the analysis of simulation data, we confirmed that the appropriate monitoring of a database could be achieved and the predictive accuracy of adaptive soft sensor models could increase by using the proposed DMI.

keywords: chemometrics, process control, soft sensor, database, monitoring

* funatsu@chemsys.t.u-tokyo.ac.jp