3A3-5

データマイニングを用いたプラントセンサデータの異常発見

Data Mining for Plant Sensor Data Fault Detection

久保田和人　森田千絵　波田野寿昭　仲瀬明彦　渡辺経夫
Kazuto Kubota　Chie Morita　Hisaaki Hatano　Akihiko Nakase　Tsuneo Watanabe

岩本撤也　大滝裕樹　大森和則　大谷圭子　河井研介
Tetsuya Iwamoto　Yuuki Ootaki　Kazunori Ohmori　Keiko Ootani　Kensuke Kawai

株式会社 東芝
TOSHIABA Corporation

Data mining technique is applied to power plant fault detection. Sensor data are monitored and compared to the value of the prediction model, and when their difference is large, the condition is judged to be abnormal. The prediction model is generated from other sensors by decision tree and regression clustering. Outline of the algorithm and its preliminary experimental results are reported.

1. はじめに

製造プラントや発電プラントでは，一度事故が起きると甚大な経済的，社会的な損失を被るため，事故を未然に防ぐことが重要な課題となる．本稿では，大量に蓄積されたセンサデータの履歴をマイニングし，異常発見の高精度化を試みた事例について報告する．

2. 従来手法と問題点

近年，プラントには大量のセンサが取り付けられており，異常の発見はセンサの値を調べることで行われる．従来手法ではセンサ値のとるべき上限と下限を定め，この範囲を逸脱した場合にプラントが異常であると判断していた．しかし，センサのとりうる値は，そのセンサが付加された設備の稼動状況に大きく依存するため，上限と下限を規定するだけでは，異常を見逃してしまう可能性がある．具体例を発電所のセンサを用いて示す．

図1は発電出力(横軸)と圧力値(縦軸)の関係をプロットしたもので，ポンプ圧力の異常を検出したいものとする．大規模な発電所では複数のポンプが設置されている場合があり，発電出力が小さいとポンプは稼動している場合としていない場合の二つの状態をとりえる．図中，グループ A が稼動状態の点の集合，グループ B が停止状態の点の集合に相当する．

ここで，P という時刻の点を考える．従来の異常発見方法では，図中右側に示した管理値の間に圧力の値が入っているため正常と見なされる．しかし，グループ A からも B からも離れており，このような圧力値と発電出力の値の組み合わせを取ることは稀で，異常である可能性が高いと言える．

3. 異常発見方式

以上のような背景をもとに，我々はデータマイニング技術を利用した異常発見システムの高精度化を試みた．異常発見の基本的な考え方は，あるセンサ(以降，ターゲットセンサと呼ぶ)の値を他のセンサ(以降，モデルセンサと呼ぶ)から予測するモデルを生成し，予測モデルと実際の値の乖離度から異常性を判定するものである．図2は，図1のデータおよびそのばらつきから計算された二つの予測モデル A, B と管理幅(図の例では3σ)を示す．この管理幅から外れたデータは異常と判定される．

連絡先：久保田和人，(株)東芝 研究開発センタ，川崎市幸区小向東芝町1，044-540-2410，kazuto.kubota@toshiba.co.jp

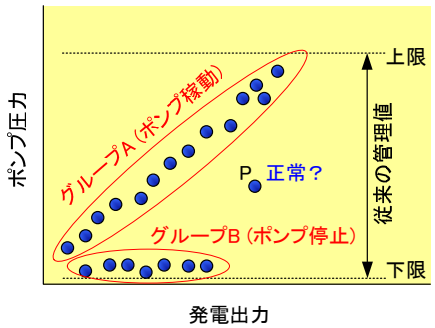

図1 発電所における圧力値と発電出力の関係

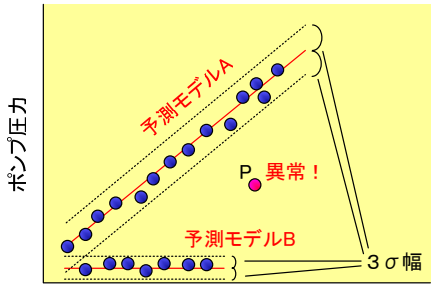

図 2 異常発見モデルおよび管理幅の算出

この方法で異常発見を行うためには二つの技術が必要となる．まず，多数のセンサの中からモデルセンサとして適切なセンサ(この例では発電出力)を見つけなければならない．次に，モデルセンサを用いて予測モデルを生成する必要があるが，この時図2のように二つのモデルを生成した方が精度の高いモデルが生成される場合は，点の分布を考慮しながら二つのモデルを生成する必要がある．これらの処理にデータマイニング技術の応用を試みる．

4. モデルセンサ選択手法

ターゲットセンサをモデルセンサから予測するためには，両者がプラント動作時に連動した動きをする必要がある．相関係数を見れば，ある程度両者が連動して動くことはわかるが，逆に相関係数が低いからといって連動した動きをしていないとは言え

ない. 図 1 の例では, グループ A, B 内でのポンプ圧力と発電出力の相関が高いが全体ではさほど高くない.

本稿では, モデルセンサ選択にデータマイニングの一手法である決定木を利用を試みる. 決定木は表形式のデータが存在したとき, ある属性の値を他の属性から予測するツリー状のルールを生成するものである[Quinlan 1993]. モデルセンサ選択の手順は, まず, 図 3 のようにポンプ圧力の変動範囲を区間に分割して各区間のデータにラベルを付加し, 続いて, このラベルを他のセンサから予測する決定木を作成する(図4). ここで, 決定木に出現するセンサはポンプ圧力と連動した動きをする可能性が高いと考え, モデルセンサの候補とする. 図4の場合, センサ X, Y, Z がモデルセンサの候補となる. なお, この段階ではモデルセンサを一つに絞らず, 複数のセンサを候補としてそれぞれ予測モデルを生成し, その中の最良ものを最終的な予測モデルとする.

なお, 候補を生成して個々に予測モデルを生成するという手順を踏んでいるのは, 全てのセンサに対して予測モデルを生成するのは計算のコストが大きいためである. センサの数が限られている場合は候補生成のステップは省略し, 全てのセンサで予測モデルの生成を試みればよい.

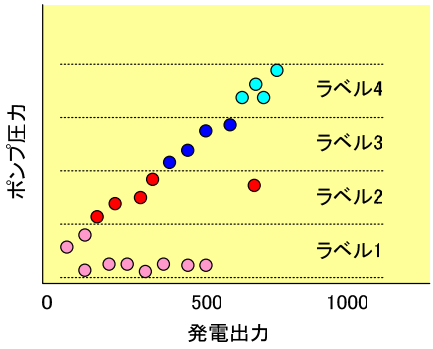

図3 ターゲットセンサの区間分割とラベル付け

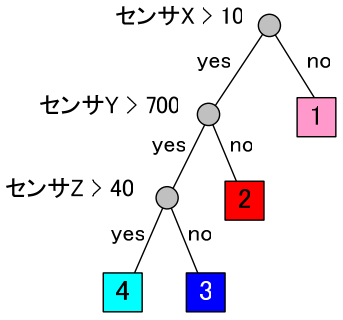

図4 モデルセンサ選択のための決定木生成

5. 予測モデル生成手法

予測モデル作成には, 重回帰クラスタリング(RC: Regression Clustering)を利用する[Zhang 2003]. 通常のクラスタリングでは点同士の距離の近いものを一つのクラスタへとまとめるが, RC では空間上の点を回帰モデルへの当てはまりを考えながらクラスタリングする. K-means 法ベースの RC アルゴリズムでは, あらかじめ K を設定し, 収束計算を行ないながらクラスタを生成するが, K の設定の問題や, モデルの初期値の問題がある. 従って, ここでは, データを再帰的に分割することで RC を行なう.

5.1 データ再帰分割による RC

図 5 に示す関数 Slice によってデータを再帰的に分割しクラスタリングを行なっていく. 動作を D が 2 次元の場合について説明する(図 6). X-Y 平面上に描かれた縦横のグリッドが Sset である. ここでは, 軸と直交する平面(直線)を候補とする. Sset から, あるsを取り出し, 入力データ D を分割してクラスタ D1s, D2s を作成する. 続いてクラスタ D1s, D2s について回帰モデル(この例では線形モデル)を求め評価値を計算する. 以上の操作を Sset 中の全ての候補について行い, 評価が最大となる smax を求め, 再帰的に Slice(D1smax), Slice(D2smax)を実行する. 評価値が閾値を超えなかった場合は処理を終了する.

```
D ← 入力点集合
Slice (D) {
    Sset ← 分割面候補
    foreach s in Sset {
        D を s で D1s, D2s へと分割
        評価値計算
    }
    評価値最大の s (= smax )を選択
    終了判定
    Slice ( D1smax ), Slice ( D2smax )
}
```

図 5 データ再帰分割による RC

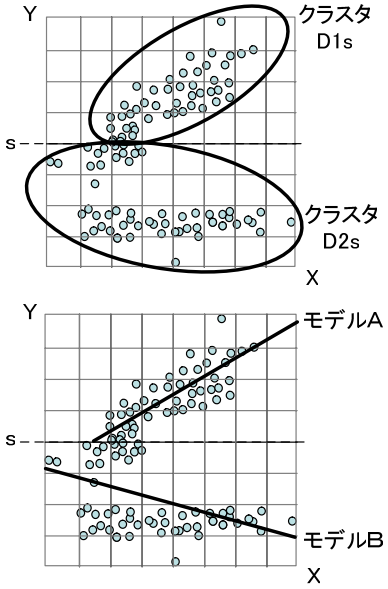

図 6 2 次元データへの適用例

5.2 評価値計算

一般に RC の評価は, モデルと点との距離の平均で計算される. しかし, 異常値検出の場合は, トータルで誤差の少ないクラスタリングよりも, 単独でも誤差が少ないクラスタを持つクラスタリングの方が望ましい場合がある. 従って, ここでは D1s, D2s について評価値を計算し, 良い方を s の評価値とする.

ここで, D1s, D2s の評価値をモデルと点との距離の平均で評価するとデータ数の少ないクラスタほど評価が高くなるという問題が生じる. 従って, クラスタ内の点の数が多いクラスタにアドバンテージを持たせるように評価値 err_{adj} を決める.

$$err_{adj} = err - \alpha \times n + \beta$$

ここで，err は点とモデルの誤差，α はデータ数 n に持たせるアドバンテージの係数，β は終了判定に用いるパラメータである．α の決め方には様々な方法が考えられるが，ここでは分割前のデータに対してモデルを生成したときの誤差 err_D を分割前のデータ数 n_D で割った値，すなわち $\alpha = err_D / n$ とした．α および β の意味を図 7 を用いて説明する．

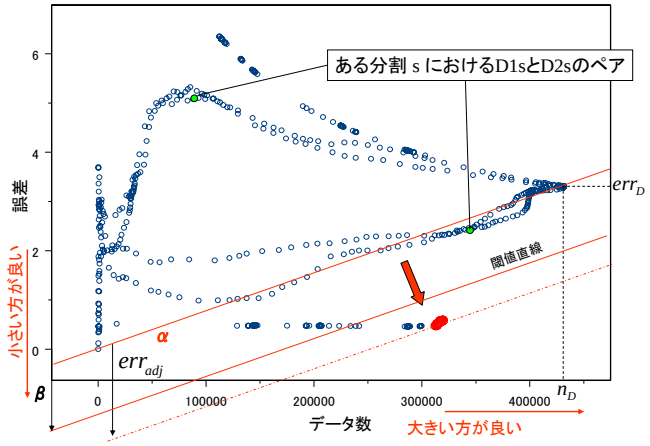

図 7 err_{adj} および α，β の直感的意味

図上の点は全ての分割面の候補を用いて D を分割した時の D1, D2 の回帰モデルに対する誤差とデータ数の関係を表したものである．一つの分割について二つの点が生成される．ここで，ある一つの点に着目した場合，誤差は小さい方がよくデータ数は大きい方が良いことになる．したがって，図中右下にある点ほど良いクラスタとなる．この右下という基準を原点を通り，傾き α の直線から最も遠い点と定める．この直線は，分割前のクラスタの回帰モデルに対する誤差 err_D が，データを分割してデータ数を減らしていくと線形に減少し，データ数が 0 になった時点でゼロになったと仮定したときの誤差とデータ数の関係を表すものである．この直線をさらに－ β 下方にずらした直線を閾値直線とし，点がこの直線の下になければ，それ以上の分割を行なわないことにする．これは，分割してもメリットが少ない場合は分割を停止するためである．

6. 実データへの適用例

6.1 モデル生成

本手法を発電所のセンサデータに対して適用した．まず，対象とするセンサを特定のブロックに属する 2521 センサとした．これらのセンサは 1 分刻みで 10 ヶ月分のデータが取得されている．この中から圧力センサ A と振動センサ A をターゲットセンサとし，残りのセンサから予測モデルを生成した．

まず，ターゲットのセンサの変動範囲を 10 区間に分割してラベル付けし，そのラベルを予測する決定木を生成した．圧力センサ A の場合は 128 種類，振動センサ A の場合は 181 種類のセンサが分類に利用された．続いて，これらのセンサと，発電出力をモデルセンサとして RC を行い，異常発見予測モデルを生成した．分割面の候補は，軸と直交し各センサの変動範囲を 100 等分するような平面を選んだ．図 8 と図 9 は圧力センサ A と振動センサ A に関して評価値が最も高かった予測モデルである．

圧力センサ A は発電出力の他にモデルセンサとして圧力センサ B が選択された．二つのクラスタが生成され，圧力センサ A が 0 の状態(図中(a))と圧力センサ B と連動して動く部分(図中(b))に分けられている．振動センサ A に関しては振動センサ B が選択された．クラスタは三つ生成され，図中(c)の部分は両者が 0 で動いていない部分，(d)と(e)は両者が連動して動いている部分であるが，発電出力の大小によって，その関係が異なっていることがわかる．

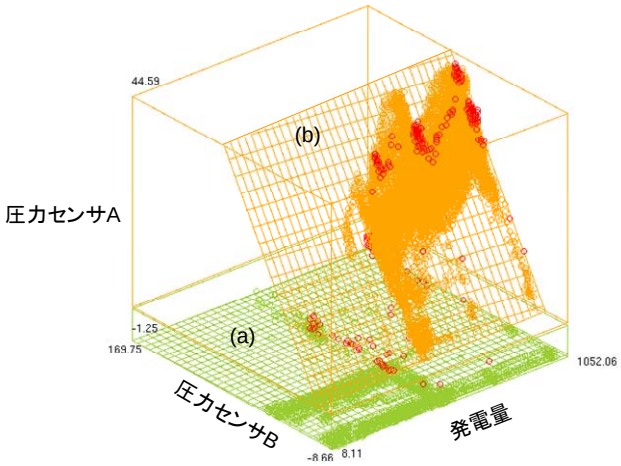

図 8 圧力センサ A の予測モデル

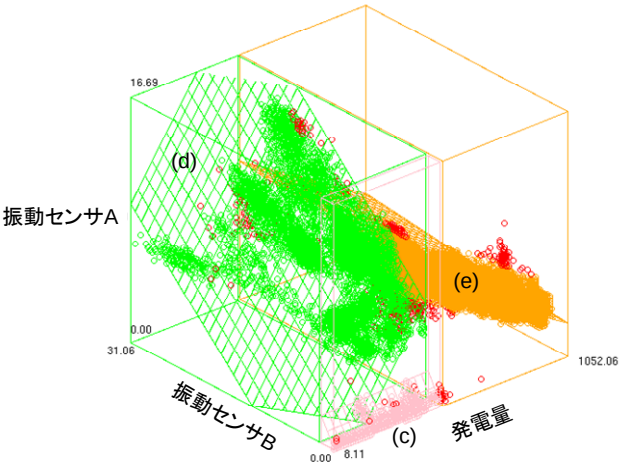

図 9 振動センサ A の予測モデル

6.2 モデル選択

二つ以上のクラスタおよび異常発見予測モデルが生成された場合，ある時刻の点がどれか一つのモデルに近ければ正常と考えても良いが，その近いモデルを基準として異常性を判定してよいかどうかはわからない．図 9 のようにモデルセンサの値とクラスタが一対一に対応する場合，対応するモデルを利用すれば基準とすべきモデルは一意に決まる．しかし，図 8 のようにモデルセンサの値に対して二つのモデルが対応する場合，モデルを選択するための指標が必要となる．

この指標作成に決定木の利用が可能である．クラスタごとに点にラベルを付け，このラベルをターゲットセンサを除いたセンサの値から予測するモデルを生成する．図 8 の例の場合，流量センサ A を条件とした木が生成された．この木がモデル選択の指針を与える(図 10)．

流量センサ A <= 33.98 (a)
流量センサ A > 33.98 (b)

図 10 モデル選択のための決定木

7. 考察

本手法は，決定木を用いたモデルセンサ選択と RC によるモデル生成を用いているが，これらはいずれもデータを再帰的に分割して処理を進めているため，両者を融合することが可能である．すなわち，決定木のデータ分割の指標である Gini Index や情報量の替わりにモデル生成および誤差の評価を行なえばよい．この方法を用いれば，クラスタ毎に異なるモデルセンサを用いたモデルが生成され異常発見の精度が向上する可能性がある．

本手法は，軸と直交する平面でデータを分割しているため，このような平面で分けられないデータを上手くクラスタリングすることができない．データを細かく分割して，再び組み上げるという方法もあるが，どこまで良いクラスタリングになるかは未知である．本手法の限界を見極め，必要に応じて K-means ベースの RC 手法を導入していく必要があると考えている．

8. おわりに

本稿では，データマイニングを用いたプラントシステムの異常発見の高精度化について報告した．異常発見モデル生成に決定木と RC を利用しているという特徴を持つ．今後は，より大規模，多種のデータを用いて手法の評価，改善を検討していく予定である．

参考文献

[Quinlan 1993] Quinlan: AI によるデータ解析，トッパン，1993.

[Zhang 2003] Bin Zhang: Regression Clustering，Proc of ICDM'03，2003.