標本時系列抽出による時系列データ特異点検出方式

中村 隆顕† 今村 誠† 平井 規郎† Michael Jones‡ Daniel Nikovski‡

†三菱電機株式会社 情報技術総合研究所 〒247-8501 神奈川県鎌倉市大船 5-1-1
‡ Mitsubishi Electric Research Laboratories 201 Broadway, 8th Floor Cambridge, MA 02139-1955
E-mail: †{Nakamura.Takaaki@dy, Imamura.Makoto@bx, Hirai.Norio@dx}.MitsubishiElectric.co.jp,
‡{mjones, nikovski}@merl.com

あらまし 時系列データからの異常検知の方式として Discord 等の方式が知られている。Discord 方式では、時系列データの部分列の近傍距離を総当たりで求め、他と外れたデータ(特異点)を検出するため、データ長の増加に従って検出時間が増加する課題がある。特に、距離を DTW 距離とした場合に、この課題は顕著である。本論文では、時系列データから抽出した標本時系列データを利用して近似的に探索することにより、高速に特異点を検出するアルゴリズムを提案する。特に、標本時系列データを効率的に抽出する方式について述べる。また、26 種類のデータに対して、提案方式の検出速度と精度を評価した結果、15 種類のデータに対して、検出精度の大きな低下を招くことなく(F 値 0.9 以上)、検出時間を削減(最大で約 1,000 分の 1)することができ、提案方式の有効性を確認した。

キーワード 時系列データ、特異点検出、Discord、標本時系列データ

1. はじめに

ビルやプラント等における、空調、電気、照明、給排水設備を始めとして､工場のラインを構成する機器、家庭用の電化製品、自動車や鉄道車両などの車載機器など、様々な機器に状態を把握するためのセンサが搭載されている。そして、これらのセンサから、時間の経過に従い発生する時系列データが収集・蓄積されている。近年、これらの時系列データを使って、機器の異常検知を行い、機器の保守を効率化するニーズが高まっている。

異常検知にも様々な手法が提案されているが、その一つに、Keogh らによる Discord 方式がある。Discord 方式では、時系列データから切り出された部分列同士の距離を総当たりで比較して、他とは最も距離が離れた部分列(特異点)を検出する。その特異点を元に、そのため、データ長の増加に従って、距離の比較回数も増加し、検出時間が長くなる課題がある。特に、距離を DTW 距離(Dynamic Time Warping)とした場合、距離の計算コストが高いため、上記の課題が顕著となる。

本論文では、時系列データの特異点検出方式において、特異点を近似的に探索することにより、高速に特異点を検出するアルゴリズムを提案する。提案方式では、近似的な探索のために、時系列データの部分列の特徴を抽出した標本時系列データを利用する。そこで、本論文では、特に、効率的な標本時系列データの抽出アルゴリズムを中心に示す。そして、26 種類の多様なデータを対象とした評価実験の結果、長い周期で類似パターンが繰り返すデータなど、15 種類のデータに対して、検出時間と検出精度で、提案方式の有効性を確認した。また、評価結果に基づき、提案方式の有効性とデータの特性について考察する。

本論文の構成は、下記の通りである。第 2 章では、時系列データの特異点検出、従来技術について述べ、本論文で扱う課題を示す。第 3 章では、第 2 章の課題を解決するための提案方式を示す。第 4 章では、速度性能、検出精度の評価結果を示す。5 章では、まとめと今後の課題を示す。

2. 課題

2.1. 時系列データ特異点検出

時系列データの特異点を検出する方法として、正常な運転データを訓練時系列データとして用いて、与えられたテスト時系列データの特異点を検知する方法がある。その方法として典型的なものは最近傍法である[1]。最近傍法では、いつものデータ挙動からの外れ度合が大きいものを特異点として抽出する。具体的には与えられたテストデータと最も近い訓練データを検索して、最も近い訓練データとの距離が大きいテストデータを特異点と判定する。しかし、時系列データの場合は、データが時点毎の値の列になっているので、テストデータと訓練データとの距離をどのように定義して良いかは自明ではない。なぜなら、Keogh[2]らが明らかにしたように、時系列データの時間を次元とするベクトルとして扱った単純な外れ値検出では、データを時間方向にずらしただけで全く違った結果になるため、二つの時系列データを比較する際には，開始時刻を合わせないと無意味になるからである。

2.2. 従来技術: Discord

Keogh らは、上記の問題を解決する高速な手法として、Discord[3][4]を提案した。Discord では、時系列データの通常から外れた部分列を求める問題を､「時系列データとウインドウサイズ w が与えられたとき、時系列データ上を、ウインドウをスライドさせることによ

り切り出される長さ w の部分列毎に、自身の部分列以外の長さ w のすべての部分列との距離の最小値が最も大きいものを求める問題」として定式化している。ここで距離は、ユークリッド距離あるいは、ユークリッド距離に対して時間方向へのずらしを許容した DTW 距離[5]を用いる。

Keogh らのオリジナルの定義では，テストデータ S と訓練データ T を分けずに、テストデータ S のみを用い、S からスライドウインドウにより切り出された部分列 S_i と、S_i 以外の S の部分列との最近傍距離が最大となる S_i を Discord 定義している。本論文では，現実の異常検知アプリケーションの問題設定に適合するように、S と T を分けた。例えば、現実のアプリケーションでは、T は機器が正常に動作していた時の時系列データ、S は直近の時系列データに対応する。このとき、テストデータの部分列 S_i と T の部分列との最近傍距離を特異度スコアと呼ぶこととする。異常検知アプリケーションでは、この特異度スコアが大きい部分列を特異点と見なす。

2.3. 本論文で扱う課題

Fig 1 に、Discord 方式のアルゴリズムの概要を示す。

```
入力:  T:訓練時系列データ、S:テスト時系列データ、
       w:ウインドウサイズ
出力:  A:異常度スコア
1   A = [];
2   for i = 1 to ||S||-w+1:
3      min = INF;
4      for j = 1 to ||T||-w+1:
5         d = lb_dtw(T(j,w),S(i,w), min);
6         if d < min:
7            min = d; idx = j;
8         end   // if
9      end   // for j
10     A[i] = (min,idx);
11  end   // for i
```

Fig 1　Discord 方式アルゴリズム(概要)

Discord 方式は、訓練時系列データとテスト時系列のデータのそれぞれから、スライドウインドウによって切り出された部分列の全組み合わせに対して、O(nm)回距離を比較する。特に、距離を DTW 距離とした場合は計算コストが高いため、Keogh らは、S と T の部分列間の距離の計算(Fig 1 の lb_dtw)において、これまでの距離の最小値(Fig 1 の min)を超えると判断できた時点で、距離の計算を打ち切ることにより、総当たりによる距離の計算回数を削減して、高速化する手法を提案している[4]。この、打ち切りのための閾値として、様々な DTW 距離の下界値が提案されている[4][6]。しかし、この手法でも、T の全ての部分列を探索することには変わりがなく、T が長大になるに従い、距離の比較回数も増加し、結果として特異点検出のための実行時間が長くなるという課題がある。

後述する noisy sine データ(訓練／テスト時系列データ長 10,000)を対象として、ウインドウサイズを 300 とした検出実験を行ったところ、単純なユークリッド距離による総当たり方式では、実行時間が 30 秒弱であった。また、DTW 距離では、5,000 秒以上の時間を要した。なお、DTW 距離算出時の Warping バンド幅は、ウインドウサイズの 1/10 の 30 とした。

3. 提案方式

3.1. 基本方式

本論文で提案する特異点検出方式は近似解法である。予め、訓練時系列データの中から選択された部分時系列データである標本時系列データの集合と、テスト時系列データの部分時系列データとの近傍距離に基づき、特異点を検出することにより、実行時間の削減を図る。標本時系列データの集合を E とすると、距離の計算回数は O(|E|m)となり、|E|が n に対して十分小さい場合は、総当たり方式と比較しての高速化が期待できる。この、標本時系列データは、予め与えられた許容近似誤差 ε の範囲に部分時系列データの集合毎に選択する。この標本時系列データ集合の選択のオーバーヘッドを小さくしつつ検出精度を維持するためには、

- 訓練時系列データの特徴を無駄無く、且つ、漏れなく抽出していること。
- 抽出のための計算量が小さいこと。

が求められる。そこで、本論文では、以下に示す時系列データの性質に着目する。

【性質 1】連続性：温度などの物理現象の計測量の多くは、連続的に変化する。また、機械的に制御された機器の計測量の場合は、値が急激に変化しない様に制御されているものも多く存在する。そのため、時系列データからスライドウインドウにより切り出された隣接した部分時系列データ間の距離は極めて近いことが多い。

【性質 2】類似パターンの頻出：機器の時系列データには、人間の活動、気象現象に強い相関を持つものがあり、それらの時系列データは、1 日、1 週間、1 年などの単位で周期性を持つ。また、機器には、プログラムに従って反復動作するものがある。これらの時系列データでは、時間的には離れているものの、類似したパターンが繰り返し現れることがある。

3.2. 定義

訓練時系列データと、テスト時系列データが与えられた場合に、Discord のアイデアに基づいた特異点検出処理を高速に実行することを目的とする。初めに、関連する定義を与える。

【定義 1】時系列：T

時系列データ $T=t_1,\ldots,t_n$ は、n 個の順序付けられた実数の列である。t_i を時点 i の値($1 \leqq i \leqq n$)、n を時系列データの長さと呼ぶ。T(i)で、T の i 番目の値 t_i を指し示すものとする。

【定義 2】部分時系列データ：T(i,w)、T_i

T を長さ n の時系列データとする。T から抽出された長さ $w \leqq n$ の連続する値の列 $T(i,w)=t_i,t_{i+1},\ldots,t_{i+w-1}$($1 \leqq i \leqq n-w+1$)を、T の部分時系列と呼ぶ。特に、部分時系列データ長 w が明らかな場合は、T_i とも表記する。

【定義 3】時系列データの平均：$\overline{S}$

S を長さ m の時系列データまたは部分時系列データの値の平均$\overline{S}$を、以下の式で定義する。

$$\overline{S} = \frac{1}{m}\sum_{j=1}^{m} S(i)$$

【定義 4】部分時系列データの標準偏差：std(S)

S の値の標準偏差 std(S)を、以下の式で定義する。

$$\mathrm{std}(S) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\overline{S} - S(i)\right)^2} = \sqrt{\overline{S^2} - \overline{S}^2}$$

【定義 5】時系列データ集合の重心：cent(U)

k 個の長さ n の時系列データの集合 $U=\{T_1,T_2,\ldots,T_k\}$とする。このとき、U の重心 cent(U)を以下で定義する。

$$\mathrm{cent}(U) = c_1, c_2, \ldots, c_n$$

$$c_i = \frac{1}{k}\sum_{j=1}^{k} T_j(i), \quad \text{但し } 1 \leq i \leq n$$

【定義 6】時系列データ間の距離下界：lb_dist(T,S)

長さが同じ時系列データ T、S と、時系列データ間の距離 dist が与えられているとする。この時、

$$dist(T,S) \geq \mathrm{lb\_dist}(T,S)$$

を満たす値 lb_dist(T,S)を、時系列データ間の下限距離と呼ぶこととする。時系列データ間の下限距離の算出方法については、3.3 にて示す。

3.3. 時系列データ間の距離下界

本論文では、時系列データ間の距離 dist を、ユークリッド距離を求める関数とする。時系列データ T、S の距離に対して、以下の二つの不等式が成立する。第一を「平均下界」($\mathrm{lb\_dist_M}$)と呼び、時系列データの平均から求めることができる。証明は、Appendix A.に示す。

$$\mathrm{dist}(T,S) \geq \sqrt{w} \cdot \left|\overline{T} - \overline{S}\right|$$

第二の不等式を「平均・偏差下界」($\mathrm{lb\_dist_{MS}}$)と呼び、時系列データの平均と標準偏差から求めることができる。証明は、Appendix B.を参照のこと。

$$\mathrm{dist(T,S)} \geq \sqrt{w} \cdot \sqrt{\left(\overline{T} - \overline{S}\right)^2 + \left(Std(T) - Std(S)\right)^2}$$

3.4. 標本時系列データ抽出アルゴリズム

3.4.1. 全体アルゴリズム

提案方式では、近似的に特異点を検出するため、予め、許容可能な近似誤差を与える。距離がこの許容近似誤差以下の部分時系列データの集合から、一つの標本時系列データを抽出する。

本論文で提案する特異点検出アルゴリズムの全体の流れを Fig 2 に示す。

```
入力:  T:訓練時系列データ、S:テスト時系列データ、
       w:ウインドウサイズ、 ε :許容近似誤差
出力:  A:異常度スコア
1  F=InitializeExemplarSet(T,w, ε );
2  E=MergeExemplarSet(F, ε );
3  A=CalcAnomalyScore(E,S,w);
```

Fig 2　全体アルゴリズム

本アルゴリズムでは、初めに標本時系列データ集合 E を生成し(1～2 行目)、次に異常度スコア A を算出する(3 行目)。

3.4.2. 標本セグメント集合の生成

標本時系列データ集合生成の基本的な考え方は、初めに時間的に近接した部分時系列データからなる初期標本時系列データ集合 F を生成する(Fig 2 の 1 行目)。次に、今度は時間的に離れた初期標本セグメント集合を結合して標本時系列データ集合 E を得る(同 2 行目)というものである。初期標本時系列データの生成アルゴリズムを、Fig 3 に示す。

```
入力:  T:訓練時系列データ、w:ウインドウサイズ、
       ε :許容近似誤差
出力:  F:初期標本時系列データ集合
1   F={};
2   for i = 1 to ||T||-w+1:
3     一時部分時系列データ集合 U = {T_i};
4     for j = i+1 to n-w+1:
5       if dist(T_i,T_j) ≦ ε /2:
6         U = U + T_j; // T_j を U に追加
7       else:
8         U の重心 C = cent(U);
9         F = F + (C,C̄,std(C),U);
10                // C、C の平均値、標準偏差と U の
11                // 組を F に追加
12        break;  // j の for ループから抜ける
13      end  // if
14    end  // for j
15    i = j;
16  end  // for i
```

Fig 3　初期標本時系列データ集合生成アルゴリズム

ここでは、時系列データの性質 1:連続性を利用し、スライドウインドウにより切り出された連続した部分時系列データの中で、距離が ε /2 以下のものを、初期標本時系列データとして抽出し、その重心を求める(9 行目)。抽出した初期標本時系列データと、その平均、標準偏差と、そこに含まれる部分時系列データの集合

の組の集合を出力する(10 行目)。

Fig 4 には、初期標本時系列データ集合の結合アルゴリズムを示す。初期標本時系列データは、重心から距離 ε /2 以下の部分時系列データの集合である。そこで、ある重心から、距離が ε /2 以下となる初期標本時系列データを集めることで、ある重心からの距離が ε 以下の部分時系列データの集合を抽出することができる。この集合の重心を、一つの標本時系列データとする。ここでは、時系列データの性質 2:類似パターンの頻出に示した、時間的に離れて存在する初期標本時系列データを、効率よく結合する手段を示している。

```
入力:  F={f1,...,fg}:初期標本時系列データ集合
       w:ウインドウサイズ、ε:許容近似誤差
出力:  E:標本時系列データ集合
1   F の要素を C̄ の昇順でソート
2   E={};
3   for i = 1 to |F|: // C̄ の昇順で F から取り出す
4     一時部分時系列データ集合 U = {fi};
5     F = F – fi;
6     for j = i+1 to |F|:
7       if lb_distM(fi,fj) ≦ ε /2:
8         if lb_distMS(fi,fj) ≦ ε /2
9               AND dist(fi,fj) ≦ ε /2:
10          U = U + fj, F = F – fj;
11        end  // if d
12      else:
13        break;  // j の for ループから抜ける
14      end  // if lb_dist
15    end  // for j
16    C = cent(U);
17    E = E + (C,C̄, std(C), U);
18    i = j;
19  end  // for i
```

Fig 4　初期標本時系列データ集合結合アルゴリズム

平均下界より、二つの時系列データ間の距離は、それらの平均の差に正の相関があることが分かる。そこで、初めに初期標本時系列データ集合を平均値の昇順でソートする(1 行目)ことにより、平均値が近い標本時系列データから順に探索することができる。また、平均下界が、距離の閾値を超えた時点で、それ以上は結合できないと判定することができる(8 行目)。また、平均・偏差下界により距離の閾値を超えると判定できた場合は、距離 dist を計算する必要が無い(9 行目)。すなわち、整列された初期標本時系列データ集合を一方向に探索するだけで、標本時系列データ集合を得ることができる。

3.4.3.　異常度スコアの算出

異常度スコアの算出については、Fig 5 に基本的なアルゴリズムのみを示す。

ここでも、標本時系列データの平均でソートすることにより、探索の範囲を効率化する(1 行目)。テスト時系列データの部分列毎に(3 から 13 行目)、E の中から平均値が最も近い標本時系列データ nn を抽出する(4 行目)。6～11 行目では、テスト時系列データの部分列と標本セグメントとの距離を計算するが、ここで計算する距離は、Discord 方式による DTW 距離であっても、ユークリッド距離であってもよい。

```
入力:  E={e1,...,eh}:標本時系列データ集合、
       S:テスト時系列データ、w:ウインドウサイズ
出力:  F:初期標本時系列データ集合
1   E の要素を C̄ の昇順でソート
2   A=[];
3   for i = 1 to ||S||-w+1:
4     nn = NN_search(E, S̄i);
5     min = dist(enn,Si), idx = min;
6     while if exists ej ∈ E:
7       d = dist(ej,Si);
8       if d < min:
9         min = d, idx = j;
10      end  // if
11    end  // while
12    A[i] = (min, idx);
13  end  // for i
```

Fig 5　異常度スコアの算出アルゴリズム

Discord 方式の場合は、距離の最小値 min が小さい値から始められるため、DTW 距離の下界値による打ち切り効果の向上が期待できる。また、ユークリッド距離を採用する場合は、nn と平均値が近い標本時系列データから順に距離の比較を行い、距離の平均下界が、最小値 min を超えた時点で、探索を打ち切って良い。また、ここでも距離の計算に、平均・偏差下界が利用できる。

4. 評価

4.1. 評価方針

本章では、提案方式の評価について述べる。評価は、高速化した Discord 方式[4]を比較対象とした。また、提案方式の距離の算出方法は、標本時系列データの生成で使用する距離をユークリッド距離、異常度スコアの算出には DTW 距離を使用した。

検出速度評価では、Discord 方式と提案方式の実行時間を比較した。検出精度評価では、Discord 方式による異常度スコアが 3σ よりも大きいテスト時系列データの部分列を正解とした。そして、提案方式の異常度スコア 3σ より大きい部分列を特異点としたときの、F 値を検出精度の指標とした。今回の評価では、提案方式の実行時間が Discord 方式の 1/10、F 値が 0.9 以上を有効性の指標とする。

提案方式の検出時間と検出精度は、入力とするデータと許容近似誤差 ε に依存するはずである。そして、この検出時間と検出精度は、トレードオフの関係にあると予想できる。そこで、様々なデータと ε に、提案方式の検出時間、検出精度のデータと、 ε 依存性を確認する。

評価用データ：評価用データには、Keogh らが公開している時系列データセット[8]と、類似パターンが頻出するデータの典型例としてノイズを含む正弦波(noisy sine)データ[7]を利用した。Table 1 にその一覧を示す。Table 1 において、ウインドウサイズは、評価 1～3 において、部分列を切り出すためのウインドウサイズである。許容近似誤差は、評価で使用した値の基準値である。ここに示した値を中心に、1/2～3 倍程度の範囲で 5 段階変化させて測定を行った。DTW 距離計算時の Warping バンド幅(Sakoe-Chiba Band)は、ウインドウサイズの 1/10 とした。

Table 1　評価用データ

データ名称	データ長		ウインドウサイズ	許容近似誤差
	訓練	テスト		
anngun_xcoord	5,625	5,625	170	750
anngun_ycoord	5,625	5,625	170	500
ARMA	10,000	100,000	100	18
chfdbchf13_2	1,875	1,875	160	2.5
chfdbchf13_3	1,875	1,875	160	4.5
chfdbchf15_1	7,500	7,500	160	6
chfdbchf15_2	7,500	7,500	160	6
ltstdb20221_2	1,875	1,875	160	5.5
ltstdb20221_3	1,875	1,875	160	4
ltstdb20321_2	1,875	1,875	200	9
ltstdb20321_3	1,875	1,875	200	1.2
mitdb100_2	2,700	2,700	300	4
mitdb100_3	2,700	2,700	300	2.5
mitdbx108_2	5,000	5,000	400	1.7
mitdbx108_3	5,000	5,000	400	4
nprs44	2,000	4,500	100	195
power_data1	11,000	15,000	700	4000
power_data2	11,000	9,040	700	4000
qtdbsel102_2	22,500	22,500	200	1.5
qtdbsel102_3	22,500	22,500	200	5
qtdbsele0606_2	700	2,300	70	4
qtdbsele0606_3	700	2,300	70	1.2
stdb308_2	2,400	3,000	400	4
stdb308_3	2,400	3,000	400	3.5
TEK	5,901	9,099	256	12
noisy sine	10,000	10,000	300	7

評価環境：評価環境のハードウェア、ソフトウェア構成は以下の通りとした。

Table 2　評価環境

CPU	Intel Core i3 550、3.20GHz、2 Core
メモリ	4GB
OS	Windows Server 2008 R2 Standard SP1
実装言語	Visual C++ 2012

4.2. 測定結果と考察

Fig 6 と Fig 7 に、提案方式と Discord 方式の検出時間の比(Search Time ratio)と、検出精度として F 値(F measure)を測定した結果を示す。Fig 6 には F 値の最小値が 0.8 以上のデータ、Fig 7 にはそれ以外のデータを示す。

Fig 6 及び Fig 7 の横軸は、右に位置する程、Discord 方式と検出能力が近いことを表す。縦軸は、Discord 方式の検出時間を、提案方式の検出時間で除算した値で、上に位置する程、提案方式の検出時間が速いことを表す。また、10^0 に示した赤の水平線に近い程、提案方式と Discord 方式の検出時間が近いことを意味する。そして、各系列は、マーカーが下方に位置する程、許容近似誤差が小さい測定結果である。

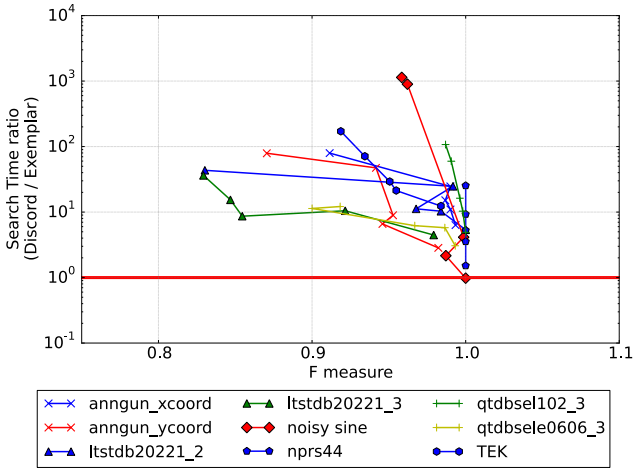

Fig 6　検出速度・検出精度測定結果(精度：高)

Fig 6 の結果より、許容近似誤差をいずれのデータも、検出時間 1/10、F 値 0.9 以上を満たす。特に、noisy sine データにおいては、最大の許容近似誤差において、提案方式の検出時間が約 1,000 分の 1、F 値も 0.95 となった。よって、これらのデータに対しては、提案方式の有効性を確認できたといえる。他のデータについても、許容近似誤差を適切に選択することにより、上記の条件を満たすことが確認できた。

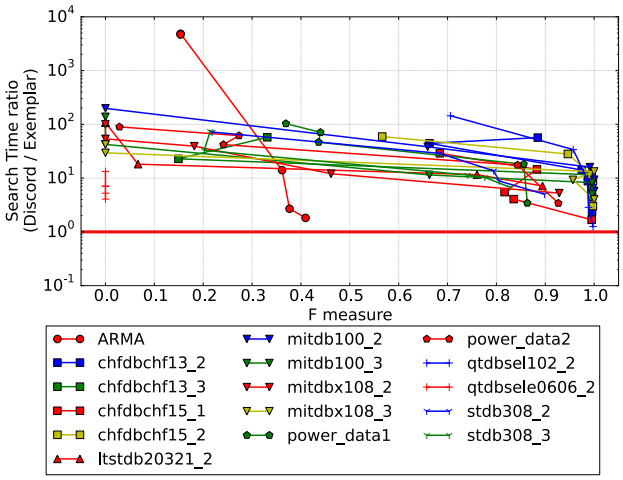

Fig 7　検出速度・検出精度測定結果(精度：低)

Fig 7 の結果では、許容近似誤差を大きくすると、F 値が 0.8 未満に低下する。ただし、chfdbchf13_2、chfdbchf13_3、chfdbchf15_2、mitdb100_2、mitdbx108_3、qtdbsel102_2 では、許容近似誤差を適切に選ぶことに

より、検出時間 1/10、F 値 0.9 以上を満たすことができる。

一方、それ以外の 10 種類のデータでは、有効性が確認できなかった。特に、ARMA、qtdbsele0606_2 では、大幅に検出精度が低下する結果となった。

Fig 8、**Fig 9** はそれぞれ、ARMA データの訓練時系列データ、異常度スコアのヒストグラムである。**Fig 9** で、異常度スコアの値は平均と 3σ で正規化しており 1.0 を超えたら特異点とみなす。

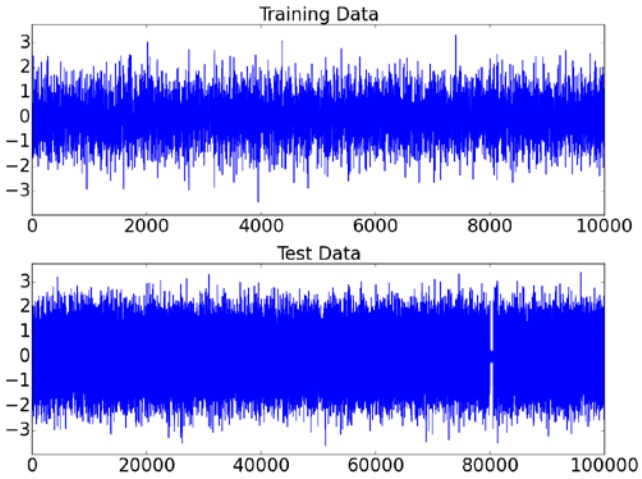

Fig 8　ARMA：訓練／テスト時系列データ

Fig 8 によると、ARMA データは、不規則に振動するノイズ様のデータであった。また、**Fig 9** によると、異常度スコアが、平均値～3σ の範囲を中心に分布している。これは、時系列データが短周期で不規則に振動しているために、訓練時系列データとテスト時系列データの部分列間の距離が近い値に集中しているものと考えられる。その結果、許容近似誤差の変化による異常度スコアの変化に従って、テスト時系列データの 80,000 付近の大きな外れは検出できるものの、1.0 に近い異常度スコアの誤差のために、過剰検出や検出漏れが多く発生、検出精度が高くなると考えられる。

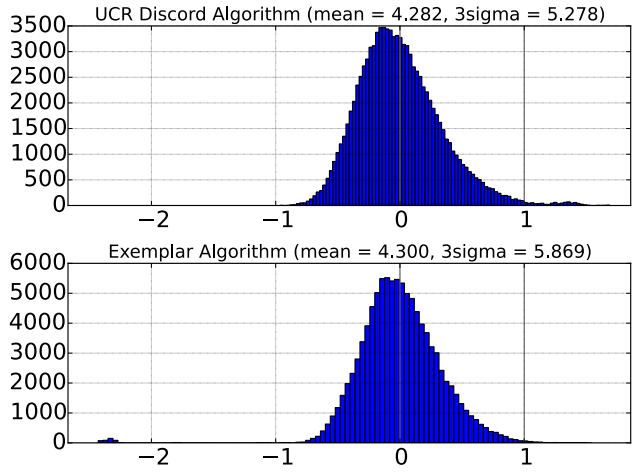

Fig 9　ARMA：異常度スコア

Fig 10 は、ARMA データにおける、適合率、再現率の推移を示したグラフであるが、特に再現率が低いことが確認できる。

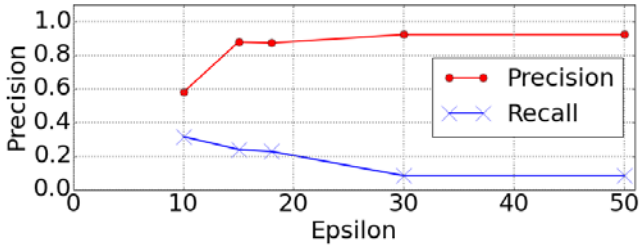

Fig 10　ARMA：検出精度

今回の評価は、ウインドウサイズを固定で行ったが、ウインドウサイズを適切に設定することにより、検出精度が向上する可能性もある。検出時間、検出精度のウインドウサイズ依存性の評価は、今後の課題である。さらには、データ毎に、最適な許容近似誤差と、ウインドウサイズを決定する方式の確立も今後の課題である。

5. おわりに

本論文では、訓練時系列データの、ある許容近似誤差の範囲で類似した部分列の集合を代表する標本時系列データ集合を用いて、テスト時系列データの特異点を近似的に検出する方式を提案した。提案方式では、標本時系列データの平均を用いて、効率的な順序で類似した標本時系列データを探索する方式と、時系列データ間の距離下界を用いて、探索範囲を枝刈りする方式についても示した。

また、26 種類の多様なデータを対象とした、検出時間と検出精度の評価結果についても示した。評価の結果、15 種類のデータに対しては、許容近似誤差を適切に設定することにより、検出精度の低下を招くことなく、検出時間を削減(最大で約 1,000 分の 1)することができ、本提案方式の有効性を確認できたといえる。ただし、残りのデータでは、近似により検出精度が低下する結果となった。有効性が確認できなかったデータの一種に、短周期で不規則に振動するデータがあることを確認した。

今後は、データの特性毎に最適な許容近似誤差とウインドウサイズを決定する方式を確立することが課題である。

参　考　文　献

[1] CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 2009, 41.3: 15.

[2] KEOGH, Eamonn; LIN, Jessica. Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowledge and information systems, 2005, 8.2: 154-177.

[3] KEOGH, Eamonn; LIN, Jessica; FU, Ada. Hot sax: Efficiently finding the most unusual time series

subsequence. In: Data mining, fifth IEEE international conference on. IEEE, 2005. p. 8 pp..

[4] RAKTHANMANON, Thanawin, et al. Searching and mining trillions of time series subsequences under dynamic time warping. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012. p. 262-270.

[5] KEOGH, Eamonn; PAZZANI, Michael. Dynamic time warping with higher order features. In: Proceedings of the 2001 SIAM Intl. Conf. on Data Mining. 2001.

[6] 大桃論, et al. タイムワーピングに基づく時系列データの類似検索: 次元縮小による効率化. DBSJ Letters, 2005, 4.1: 1-4.

[7] JONES, Michael, et al. Anomaly Detection in Real-Valued Multidimensional Time Series. 2014.

[8] Keogh, E., Zhu, Q., Hu, B., Hao. Y., Xi, X., Wei, L. & Ratanamahatana, C. A. (2011). The UCR Time Series Classification/Clustering Homepage: www.cs.ucr.edu/~eamonn/time_series_data/

Appendix A.

平均下界

$$\mathrm{dist}(T,S) \geq \sqrt{n}\left|\overline{T}-\overline{S}\right| \quad \text{(式 A1)}$$

を示す。

【証明】

長さが等しい二つの時系列データ $T=t_1,t_2,\ldots,t_n$、$S=s_1,s_2,\ldots,s_n$ が与えられ得た時、コーシー・シュワルツの不等式

$$\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \sum_{i=1}^{n} {x_i}^2 \cdot \sum_{i=1}^{n} {y_i}^2 \quad \text{(式 A2)}$$

より、(式 A1)が導出できることを示す。

(式 A2)において、$x_i = (t_i - s_i)$、$y_i = 1$と置く。

$$\left(\sum_{i=1}^{n}(t_i - s_i)\cdot 1\right)^2 \leq \sum_{i=1}^{n}(t_i - s_i)^2 \cdot \sum_{i=1}^{n} 1^2 \quad \text{(式 A3)}$$

ここで、

$$\text{左辺} = \left(\sum_{i=1}^{n} t_i - \sum_{i=1}^{n} s_i\right)^2 \quad \text{(式 A4)}$$

であり、$\sum_{i=1}^{n} t_i = n\cdot\overline{T}$なので、

$$\text{左辺} = \left(n\cdot\overline{T} - n\cdot\overline{S}\right)^2 \quad \text{(式 A5)}$$

を得る。また、右辺は、

$$\text{右辺} = \left(dist(T,S)\right)^2 \cdot n \quad \text{(式 A6)}$$

なので、

$$n\cdot\left(\overline{T}-\overline{S}\right)^2 \leq \left(dist(T,S)\right)^2 \quad \text{(式 A7)}$$

となり、(式 A1)を導出することができる。

Appendix B.

平均・偏差下界

$$\mathrm{dist(T,S)} \geq n\cdot\sqrt{\left(\overline{T}-\overline{S}\right)^2 + \left(Std(T)-Std(S)\right)^2} \quad \text{(式 B1)}$$

を示す。

【証明】

(式 B1)の左辺 ^{2}－右辺 ^{2}≧0 となることを示す。

まず左辺は、

$$\begin{aligned} \text{左辺}^2 &= \sum_1^n (t_i - s_i)^2 \\ &= \sum_1^n {t_i}^2 + \sum_1^n {s_i}^2 - 2\sum_1^n t_i s_i \\ &= n\overline{T^2} + n\overline{S^2} - 2n\overline{T\cdot S} \end{aligned} \quad \text{(式 B2)}$$

となる。次に右辺は、

$$\begin{aligned} \text{右辺}^2 &= n\cdot\left\{\left(\overline{T}-\overline{S}\right)^2 + \left(Std(T)-Std(S)\right)^2\right\} \\ &= n\cdot\left\{\left(\overline{T}^2 + \overline{S}^2 - 2\cdot\overline{T}\cdot\overline{S}\right)\right. \\ &\quad \left. + \left(Std(T)^2 + Std(S)^2 - 2\cdot Std(T)\cdot Std(S)\right)\right\} \end{aligned} \quad \text{(式 B3)}$$

となり、ここで、$Std(T)^2 = \overline{T^2} - \overline{T}^2$なので、

$$\begin{aligned} \text{右辺}^2 &= n\cdot\left\{\left(\overline{T}^2 + \overline{S}^2 - 2\cdot\overline{T}\cdot\overline{S}\right)\right. \\ &\quad \left. + \left(\overline{T^2} - \overline{T}^2 + \overline{S^2} - \overline{S}^2 - 2\cdot Std(T)\cdot Std(S)\right)\right\} \end{aligned} \quad \text{(式 B4)}$$

となる。よって、

$$\begin{aligned} &\text{左辺}^2 - \text{右辺}^2 \\ &= -2n\cdot\overline{T\cdot S} + 2n\cdot\overline{T}\cdot\overline{S} + 2n\cdot Std(T)\cdot Std(S) \end{aligned} \quad \text{(式 B5)}$$

となり、(式 B1)を示すためには、(式 B5)の右辺を整理した(式 B6)が成立することを示すことができればよい。

$$Std(T)\cdot Std(S) - \left(\overline{T\cdot S} - \overline{T}\cdot\overline{S}\right) \geq 0 \quad \text{(式 B6)}$$

ここで、コーシー・シュワルツの不等式(式 A2)において、$x_i = \left(t_i - \overline{T}\right)$、$y_i = \left(s_i - \overline{S}\right)$と置くと、

$$\left(\sum_{i=1}^{n}\left(t_i-\overline{T}\right)\left(s_i-\overline{S}\right)\right)^2 \leq \sum_{i=1}^{n}\left(t_i-\overline{T}\right)^2 \cdot \sum_{i=1}^{n}\left(s_i-\overline{S}\right)^2 \quad \text{(式 B7)}$$

左辺は、T と S の共分散の二乗

$$\text{左辺} = \left(\overline{T\cdot S} - \overline{T}\cdot\overline{S}\right)^2 \quad \text{(式 B8)}$$

右辺は、T の分散と S の分散の積

$$\text{右辺} = Std(T)^2 \cdot Std(S)^2 \quad \text{(式 B9)}$$

であるから、

$$\left(\overline{T\cdot S} - \overline{T}\cdot\overline{S}\right)^2 \leq Std(T)^2 \cdot Std(S)^2 \quad \text{(式 B10)}$$

となり、(式 B6)は 0 以上となる。以上より、(式 B1)が示される。