CF-003

# 外れ値を含む長期データ列に対する類似度計算法の検討

佐藤 哲 †

NHN テコラス株式会社データホテル事業本部事業戦略室
データサイエンスチーム †

## 1. はじめに

我々は，様々なセンサデータや色々な分野での電子商取引データなど，同時にデータ分析をしなければならないが実際には種類が異なりデータ列同士の比較が難しい，というデータ列群の解析に取り組んでいる．比較が難しい理由は，あるデータ列とあるデータ列ではデータ量やノイズの影響で平均も分散も異なることにより一方を基準とするともう一方のデータ値が外れ値であるという統計値が現れたり，欠損値の影響でデータ同士の次元が一致せず，従来の類似度計算手法が適用できない点などが上げられる．例えば，ノイズが含まれているデータ列同士を比較するためには平滑化等のフィルタリングが必要であり，欠損値が含まれているデータ列同士の類似度を計算するためには，欠損値を補わないとコサイン類似度等の計算手法は適用できない．

そこで本発表では，大域性・局所性などを考慮した幾何的な特徴量を抽出し，さらにデータ量に応じた動的な規格化手法を導入することで，比較が難しい外れ値を含むデータ列同士の類似度を計算し分類する手法について提案する．

## 2. 外れ値データ列を含むデータ列群

本発表で分析対象とするいわゆるビッグデータと呼ばれるデータ群は，統計的な分析が難しい．理由は，データのソースが多様であることと，ノイズや欠損値の混入が必然的であることがあげられる．データのソースが多様であるとは，例えばセンサデータが対象の場合は画像，位置情報，角度情報など単位が異なるデータを統合的に解析する必要があったり，電子商取引データが対象の場合はユーザが個人・法人が混ざっており取引規模が異なる，取引分野が消耗品と耐久品が混ざっており取引頻度や金額が異なるようなケースの混在があげられる．ノイズや欠損値の混入が必然的であることも同様であり，センサデータなら常に外的要因が変化するため，電子商取引データなら購買間隔が空くことで欠損値のように取り扱わないとならないケースが有り得る．

そのようなデータの例として，図 1 に ID={1, 4, 8} のラベルが付けられた 3 つのデータ列を示す．変動がある離散データ列であり，平均値は各 $\{5.7, 5.6, 11.6\}$，分散は $\{13.0, 160.0, 18.9\}$，回帰式は $\{y = 4.0145 + 0.0046x, y = 3.4886 + 0.00515x, y = 11.701 - 0.00017x\}$，観測データ数は $\{515, 252, 140\}$ 個である．分散の値及びグラフより，ID=4 のデータは大きな外れ値を含んでいることが分かり，突発的であることからこれを短期的な外れ値と呼ぶことにする．また回帰式より，ID=1 及び ID=4 の傾向は似ており，ID=8 が高い平均値を持ち他と外れていることが分かる．本発表では，ID=8 のようなデータ列を長期的な外れ値と呼ぶことにする．

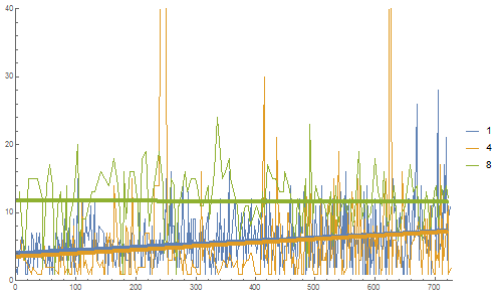


図 1: 外れ値を含むデータ列例

本発表の目的は，短期的及び長期的な外れ値を含むデータを分類することである．データを分類するためには，データ同士を比較しなければならないが，短期的及び長期的な外れ値を含むデータの場合は，局所的及び大域的な特徴量を考慮する必要がある．このような場合，局所的なデータ解析と大域的なデータ解析を統合する多重解像度解析が適用されることが一般的であるが，多重解像度解析は解像度の変化が離散的であることが多いことに加え，離散化パラメータが全てのデータに対し一律に決定されるため，性質の異なる複数のデータ列に対し適用することが難しい．従って本発表では，より連続的に解像度を切り替えながらデータ解析が可能な手法を導入する．

## 3. 位相的データ解析と順序保存符号化

データ列の局所的な特徴と大域的な特徴の両方を分析対象とするために，時系列データに対する位相的データ解析 [1] を適用する．位相的データ解析では，時系列データの時間に相当するパラメータを変化させてデータを解析するフィルトレーションと呼

A study on measuring similarity for long-term data sequence with outliers
†Tetsu R. Satoh, NHN Techous Corp.

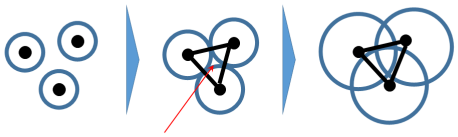

図 2: フィルトレーション例

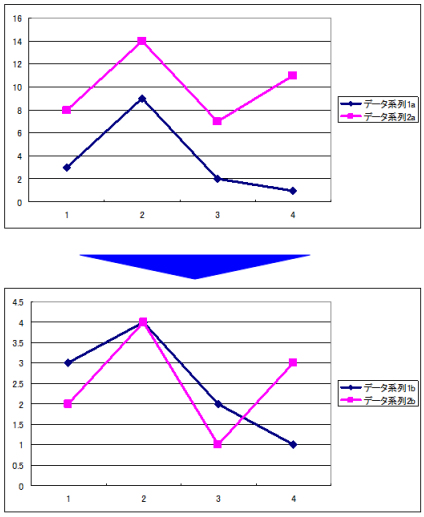


図 3: 順序保存符号化例

ばれる技術を導入することで，時間的に近い情報と遠い情報の両方を考慮した分析が可能である．つまり，位相情報だけではなく位相の変化を検出する．

フィルトレーションは，単調増加パラメータ $t_i$ に対応した単体的複体 $X_{t_i}$ の列により定義される：

$$X_{t_0} \subset X_{t_1} \subset \cdots \subset X_{t_k} \cdots \subset X_{t_n} \qquad (1)$$

ここで，単体的複体は点群 $x_i (i < m)$ の各点を中心とした半径 $r_{t_k}$ の球面を配置した時に，球面同士が交差した点を結ぶことで生成される．図 2 に，$m = 3$ の点群に対するフィルトレーションの例を示す．左から順に，0-単体である点が 3 つから始まり，1-単体である線分が 3 つ生成されている．2-単体である三角形がどのタイミングで生成されるかは，採用する幾何モデルにより異なる．また，中心の図では 3 球面が接することにより穴が発生しており，右の図では球面半径の増大により穴は消滅している．穴が発生した時刻を発生時刻，穴が消滅した時刻を消滅時刻と呼び，発生時刻を $x$ 座標，消滅時刻を $y$ 座標として 2 次元平面上に図示したものをパーシステンス図と呼ぶ．パーシステンス図では，性質上必ず発生時刻 $\leq$ 消滅時刻となるので，データの存在領域は $y = x$ という直線より上の部分となる． このように，フィルトレーションによって繋がりや穴の個数といった位相的な情報を計算できる．しかし距離の定義に依存するため，データの取りうる値の範囲が異なる複数のデータ列を比較する場合は値の規格化などの工夫が必要となる．規格化の問題に対しては，本発表では順序保存マッチング [2] の手法を採用する．順序保存マッチングはデータの規格化のための手法では無いが，データに順序保存符号化を適用することにより，データ列長に依存する一定の範囲内にデータ値が含まれるように，元のデータを変換することができる．順序保存符号化は，一般に二つのデータに対し次の性質を保つ変換のことである：

$$x_i \leq x_j \Leftrightarrow y_i \leq y_j, \forall i, j \qquad (2)$$

ここで，あるデータ $(x_1, x_2, x_3)$ に対し $(y_1, y_2, y_3)$ を $(1, 2, 3)$ とすると，$x_i < x_j$ と $y_i < y_j$ が同値であることは，データ列 $x_i$ が昇順にソートされており，その順位に対応付けることに他ならない．つまり，$x_i$ から $y_i$ に変換することにより，値の範囲をデータの個数以下の範囲に写像することができる．図 3 に，y 軸方向の値が異なる 2 つのデータ列を順序保存符号化した例を示す．値の変化の傾向は保たれたまま，地域が同じ範囲に変換されていることが分かる．

入力データを順序保存符号化したものにスライディングウィンドウを適用して一定期間のデータを抽出し，位相的データ解析対象の点群とする．例えば，連続した 2 個あるいは 3 個のデータを抽出し，2 次元空間あるいは 3 次元空間中の点データとする．そして点群のフィルトレーションからパーシステンス図を計算し，各データ列のヒストグラムを得る．最後に，ヒストグラム間の類似度を計算することで，データ列群を分類する．

## 4. 実験と考察

提案手法の有効性を検証するため，インターネットサービスで実際に収集したデータに対し適用した結果について述べる．

分析対象とするデータは，ユーザが個人・法人の両方を対象とするインターネットサービスである．従って，データ値の範囲やピーク値が発生する時期などは一律ではない．実験に用いたデータの例を図 4 に示す．分かりやすさのため，全データの中から ID=1 から ID=10 の値で識別される 10 系列のデータを対象として実験結果を説明する．横軸は 2014 年 1 月 1 日からの日数で，縦軸はアクセス数である．図より，大きな短期的な外れ値が存在することが見て取れる．

この入力データに対する処理の流れは次の通りである．

(1) 入力データを順序保存符号化することで，外れ値情報を保存しつつ値の範囲を限定する
(2) 順序保存符号化されたデータに対し一定サイズのウィンドウで複数の $y$ 座標を抽出する

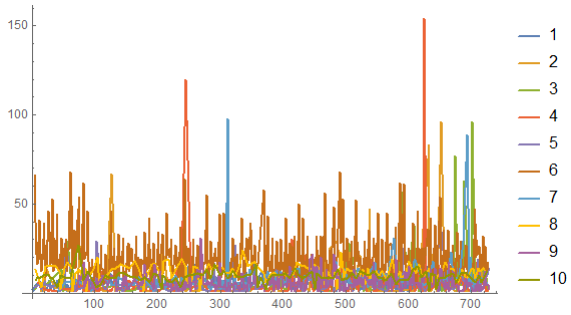


図 4: 入力データ

(3) 抽出された座標情報に対し，パーシステンス図を作成する
(4) パーシステンス図より得られるヒストグラム情報から，分布間の類似度を計算する
(5) 類似度情報を元にクラスタリングすることでデータ列群を分類する

まず，図 5 は入力データに対し値の順序保存符号化とサイズ=2 のウィンドウにより連続した 2 点のアクセス数を抽出し，2 次元平面上に図示したものである．原点付近の点密度が高いことからアクセス数が少ないデータ列の方が多いことが分かり，直線成分が現れているデータ列は散らばりが小さいことを表している．

図 6 が，図 5 のデータから作成したパーシステンス図である．位相的データ解析の計算にはライブラリ JavaPlex[1] を利用し，幾何モデルはウィットネス複体 [3] を使った．

得られたパーシステンス図を $x$ 軸方向に 10 分割してヒストグラムを生成し，各データ列間のバチャタリア係数を近似的な距離とみなし距離行列（非類似度行列）を計算し濃淡画像にて可視化した結果が図 7 であり，この距離行列を元に階層的クラスタリングを実施した結果が図 8 である．バチャタリア係数は以下の定義を用いた：

$$\rho = \sum_{k=1}^{n} \sqrt{p_k q_k} \tag{3}$$

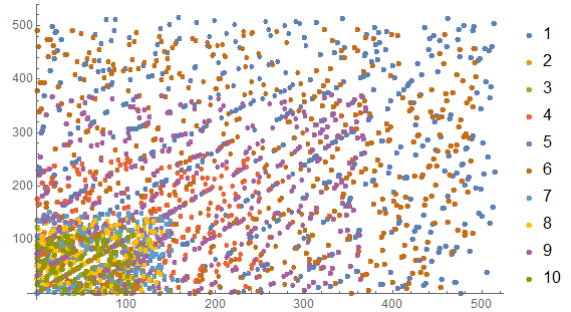


図 5: 順序保存符号化＋ウィンドウ適用

[1]https://github.com/appliedtopology/javaplex

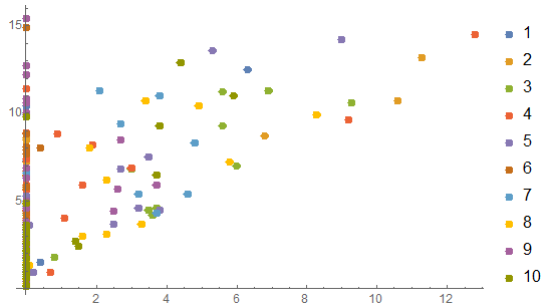


図 6: パーシステンス図

ここで $n$ はヒストグラムの区間数（本実験では 10）であり，$p_k$ 及び $q_k$ は $k$ 番目の区間の値である．なお，バチャタリア係数を計算する際には入力値を正規化しているが，順序保存符号化しても結果は変わらなかった．

ここで，どのようなクラスタリング結果が得られたかを考察するために，最も近いクラスタ要素である ID=1 及び ID=2 のデータ列と，最も遠いクラスタ要素である ID=8 の合計 3 つのデータ列を抽出して比較してみる．図 9 は 3 つのデータ列の元データ，図 10 及び図 11 はそれぞれ元データに対し順序保存符号化とサイズ=2 のウィンドウにより連続した 2 点のアクセス数を抽出した結果，及びそれに対するパーシステンス図である．

図 9 及び図 10 からは，3 つのデータ列の類似性を判断することは人間には困難である．ID=1 と比べると，距離行列上は近いと計算されている ID=2 は短期的な外れ値により類似していないように見え，最も遠いと計算されている ID=8 は元データの値が比較的小さく ID=1 に近いように見える．ところが，図 11 からは ID=8 が他と異質であることは明らかに分かる．また図 12 に示すように，短期的な外れ値の影響を軽減するために移動平均によりフィルタリングすると類似性が分かる．ID=2 は ID=1 と比べると，短期的な外れ値の影響を軽減すると値が近づき，一方で ID=8 は定常的に値が異なる長期的な外れ値となっている．また，長期的に値が増加傾向にある ID=1 及び ID=2 と比べ，ID=8 は増減の傾向は見られない．従って本手法では，短期的な外れ値の影響を受けずに長期的な外れ値という観点からクラスタリング出来ているものと考えられる．

以上の実験は Apache Hadoop/Spark クラスタ上で実施し，プログラミング言語は主に Scala-2.11(Java-1.8) を用いた．クラスタのマシンのうち，主要な計算処理を実施するワーカノードは 3 台で，スペックは CPU は Intel Xeon E5-2643，メモリ 128G バイトである．

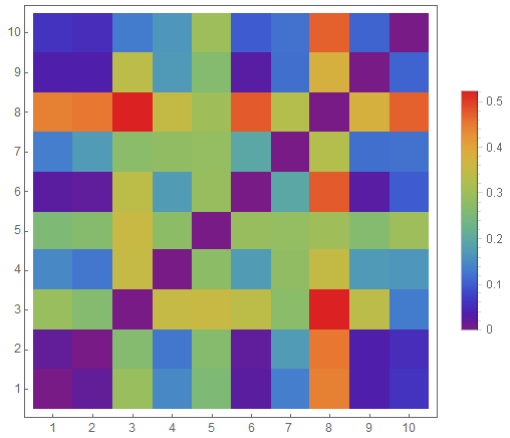

図 7: 距離行列

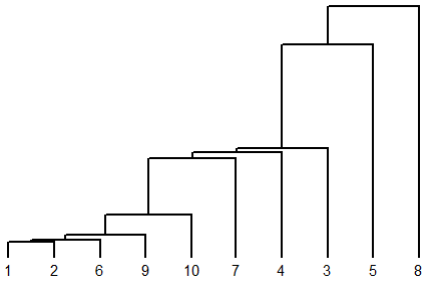

図 8: 階層的クラスタリング結果

## 5. おわりに

本発表では，外れ値や欠損値の混入が避けられないことなどが要因で互いの比較が難しいデータ列群に対し，局所的及び大域的な幾何特徴量を抽出し，データ量に応じた動的な規格化処理を行うことで，データ列同士の類似度を計算する手法を提案した．実験結果により，短期的な外れ値と長期的な外れ値の両方を考慮した類似度計算及び分類が可能であることが示唆された．

### 謝辞

元データの分析に関し，独自の分析結果や議論を通じた示唆などを提供してくれた，同データサイエンスチームの趙漢哲博士に深く感謝する．

## 参考文献

[1] L. M. Seversky, S. Davis and M. Berger, On Time-Series Topological Data Analysis: New Data and Opportunities, Conf. CVPR Workshops, pp. 1014-1022, 2016.

[2] J. Kim, P. Eades, R. Fleischer, S.-H. Hong, C. S. Iliopoulos, K. Park, S. J. Puglisi and T. Tokuyama, Order-preserving Matching, Theor. Comp. Sci., Vol. 525, pp. 68–79, 2014.

[3] V. de Silva and G. Carlsson, Topological Estimation Using Witness Complexes, Proc. 1st Eurographics Conf. Point-Based Graphics, pp. 157–166, 2004.

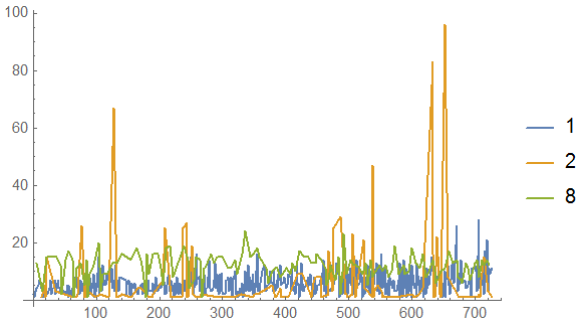

図 9: 入力データ（一部）

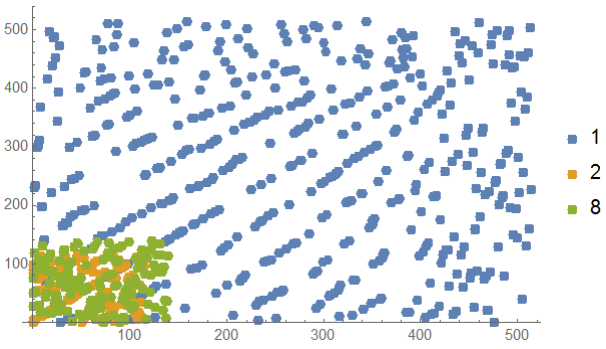

図 10: 順序保存符号化＋ウィンドウ適用（一部）

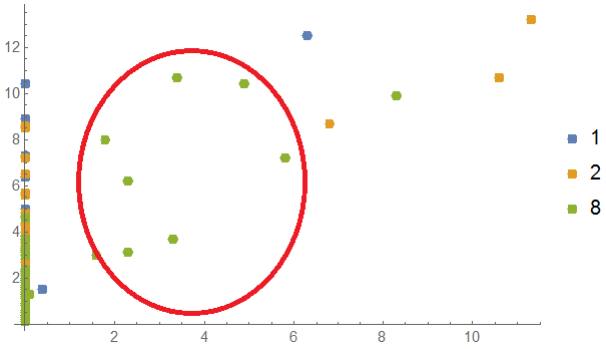

図 11: パーシステンス図（一部）

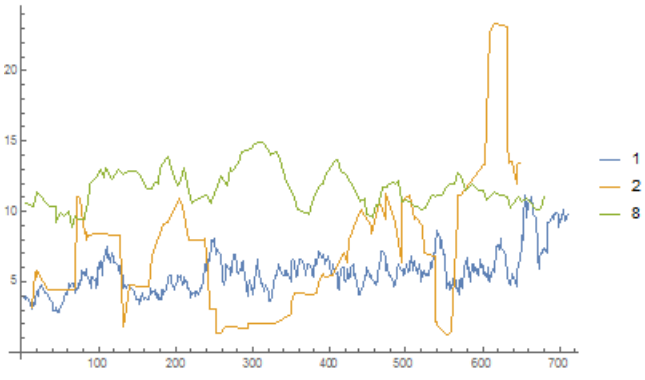

図 12: 移動平均結果