動的時間伸縮法に基づく時系列データの高速クラスタリング

LG-3

山田 悠 Yuu YAMADA　中本 和岐 Kazuki NAKAMOTO　鈴木 英之進 Einoshin SUZUKI

横浜国立大学大学院工学府物理情報工学専攻電気電子ネットワークコース
Department of Electrical and Computer Engineering, Graduate School of Engineering, Yokohama National University

1. はじめに

時系列データは，KDD が対象とする実問題において頻繁に現れる [Keogh 00]．金融における株価，医療における検査値など，その時間的変化が極めて重要な意味を持ち，そこから知識発見が求められる時系列データは多種存在する．時系列データにクラスタリングを適用することにより，このような知識が発見される可能性が高い．

本研究では，動的時間伸縮法 [Sakoe 78] の結果をボトムアップにマージして得られる平均時系列を提案し，平均時系列に関する高さ平衡木を用いて時系列データを圧縮した．その結果時系列データの高速なクラスタリングが可能となった．

2. 動的時間伸縮法

動的時間伸縮法 (DTW)[Sakoe 78] は，時系列データのペアに関する類似度計算法であり，時系列データにおける 1 点のデータをもう片方の時系列データにおける複数点のデータに対応づけられるため，時間方向の非線形な伸縮を許容できる．

2 本の時系列を．$\boldsymbol{A} = a_1, a_2, \ldots, a_I, \boldsymbol{B} = b_1, b_2, \ldots, b_J$ と表す．ただし，本論文では観測値が等間隔 (=1) の場合を扱う．$\boldsymbol{A}, \boldsymbol{B}$ の対応づけをワーピングパスと呼び，$I \times J$ 平面上の格子点 $f_k = (i_k, j_k)$ の系列で表し，$\boldsymbol{F} = f_1, f_2, \ldots, f_k, \ldots, f_K$ となる．

ワーピングパスにおいて，水平または垂直方向の線分を倍率対応，斜め方向の線分を等率対応と呼ぶ．a_{i_k} と b_{j_k} との距離を $\delta(f_k) = |a_{i_k} - b_{j_k}|$ で表す．F の評価関数 $\Delta(F)$ は式 (1) で表される．

$$\Delta(F) = \frac{1}{I+J} \sum_{k=1}^{K} \delta(f_k) w_k \tag{1}$$

この値が小さいほど，$\boldsymbol{A}, \boldsymbol{B}$ が類似していることになる．極端な伸縮を防ぐために整合窓 $r \geq |i_k - j_k|$ の条件で，$\Delta(F)$ を F に関して最小化する．ただし，w_k は f_k に関する正の重みで，$w_k = (i_k - i_{k-1}) + (j_k - j_{k-1})$ とする．また，$i_0 = j_0 = 0$.

3. 提案手法

3.1 動的時間伸縮木

動的時間伸縮木 (DTWS 木) は，時系列データを圧縮するための高さ平衡の 2 分木である．DTWS 木は，判別ノードと葉からなり，それらはそのノードを流れた例数 N とその時点での平均時系列を示す $\boldsymbol{ATW}$ の 2 要素で構成される $\boldsymbol{DTWS}$ からなり，$\boldsymbol{DTWS} = (N, \boldsymbol{ATW})$ で表される．

例を読む度に，この例を根から葉まで流し，経路上の全てのノードを更新する．判別ノードにおいては，DTW に基づき最も近いノードに例を流す．また，葉において DTW に基づく距離が帰属判定閾値 α 以内ならば，その例によって葉の $\boldsymbol{DTWS}$ を更新する．α より大きければ，この例を平均時系列ベクトルとする例数 1 の新しい葉を生成する．

$\boldsymbol{DTWS_1} = (N_1, \boldsymbol{ATW_1})$ と $\boldsymbol{DTWS_2} = (N_2, \boldsymbol{ATW_2})$ から，$\boldsymbol{DTWS} = (N_1 + N_2, \boldsymbol{ATW})$ に更新する流れを説明する．$\boldsymbol{ATW}$ は，例数 N_1 と N_2 の割合を考慮して更新される．$N_1 \geq N_2$ と仮定する．まず，式 (2) から整数 n を求める．$nearinteger$ は，最も近い整数を求める関数である．

$$n = nearinteger\left(\log_2 \frac{N_1 + N_2}{N_2}\right) \tag{2}$$

$\boldsymbol{ATW}$ は，式 (3) を $i = 1$ から $i = n$ まで繰り返すことによって求めることができる．

$$\boldsymbol{ATW_{n_i}} = avewarp(\boldsymbol{ATW_1}, \boldsymbol{ATW_{n_{i-1}}}) \tag{3}$$

ただし，$avewarp$ は 2 つの時系列から平均時系列を求める関数であり，その方法は次節で述べる．また，$\boldsymbol{ATW_{n_0}} = \boldsymbol{ATW_2}$ とする．

3.2 平均時系列

BIRCH[Zhang 96] を時系列データに適用する場合，時系列データのペアに関する距離を求めなければならない．そこで，平均時系列の作成が必要になる．

2 時系列データの各点を同時刻で対応させ，平均値を求めて新しい時系列データを作成することを考える．2 時系列データの凸部分が時間軸方向にずれている場合，結果としてできる時系列は凸部分がつぶれた形のものとなる．この問題に対し，2 時系列からワーピングパスに基づき平均時系列を生成することを提案する．この平均時系列は，時間軸方向における凸部分のずれを許容できる．

2 つの時系列 $\boldsymbol{A}, \boldsymbol{B}$ から平均時系列 $\boldsymbol{H}$ を求める方法を説明する．まず，DTW から $\boldsymbol{A}, \boldsymbol{B}$ の対応を調べるため，ワーピングパス $\boldsymbol{F}$ を求める．次に，$\boldsymbol{F}$ の全ての点の値を，平均値 $h_k = (a_{i_k} + b_{j_k})/2$ に更新する．$\boldsymbol{F}$ を指定長 K' に収縮する．収縮において，倍率対応は，等率対応の半分の割合で収縮すべきであるが，この割合で収縮すると直観に反する結果が得られる場合がある．

そこで，ワーピングパス $\boldsymbol{F}$ を対応が同じ部分毎に収縮する方法を考案した．ワーピングパス $\boldsymbol{F}$ を $L(\beta)$ 個の部分

連絡先: 山田 悠，横浜国立大学大学院工学府物理情報工学専攻電気電子ネットワークコース 鈴木研究室，〒 240-8501 横浜市保土ヶ谷区常盤台 79-5, Tel: 045-339-4135, Fax: 045-339-4148, E-mail: yuu@slab.dnj.ynu.ac.jp

$H_1, H_2, \ldots, H_{L(\beta)}$ に分け，各部分の重み w_k を求める．ただし，β は結合閾値と呼ばれユーザが指定する値であり，この値によって $L(\beta)$ が決まる．平均時系列 $\boldsymbol{H}$ は，各部分を w_k に比例した割合で長さ K' に収縮して，線形補間して等間隔の時系列を得る．

3.3 振幅伸長による平均時系列

先の平均時系列を作成する方法では，少しずつ縦方向に収縮する可能性がある．平均時系列の更新回数が多ければ，収縮の度合も大きくなってしまう．そこで，縦方向に伸ばす必要がある．$\boldsymbol{DTWS}$ 更新で，まず，$\boldsymbol{DTWS}_1$ と $\boldsymbol{DTWS}_2$ の入力から，平均振幅 $aveamp$ を求める．$\boldsymbol{ATW}_1$ と $\boldsymbol{ATW}_2$ の振幅をそれぞれ amp_1, amp_2 とすると，$aveamp$ は次式で表される．

$$aveamp = \frac{N_1 amp_1 + N_2 amp_2}{N_1 + N_2} \tag{4}$$

最終更新した平均時系列である $\boldsymbol{ATW}_{n_n}$ の振幅を amp_n とし，振幅比 $\gamma = aveamp/amp_n$ を求める．時系列データの凸部分の特徴を強調するため，凸部分だけ γ を掛け縦方向に伸ばした平均時系列を作成する．実験で使用した手話データの場合，最大，最小の絶対値の 95%以上の部分だけ γ で縦方向に伸ばした．

4. 実験評価

4.1 実験準備

オーストラリア手話データ [Bay 99] を用いて，DTWS 木による例圧縮をした場合と，例圧縮無しの場合とのクラスタリング実験で評価する．オーストラリア手話データは，手話の種類をクラスとするデータである．ここでは，クラスを隠してクラスタリングを行い，クラスタと手話の種類の一致度を正答率として表した．クラスタリングの正確性と実行時間を評価指標とした．クラスとして実験者の一人である Waleed 氏の 6 種類の手話で，x 軸に関する時系列データだけを用いた．

実験で用いた手話データは，クラスの種類が異なっていても形が似ている例がある．よって，6 種類のうち 3 種類の組み合わせで，データ圧縮せずにクラスタリングし，正しい結果が得られる 3 種類の 3 クラスデータを用いることとした．手話データは例数が少なすぎるため，時系列データの一部を左右に動かして，人工的に例数を増やした．これにより，各データ集合の例数はそれぞれ 1395 となった．

DTWS 木を用いたデータ圧縮では，DTW の窓幅率を 0.10，$K' = 50$, $\beta = 5$ とした．また，α を変化させ種々の圧縮率で比較した．

4.2 実験結果

図 1 に，“forget, innocent, lose”，図 2 に” forget, later, Norway”，図 3 に”forget, Noray, spend” のクラスタリング結果を示す．それぞれ線グラフは k-medoid によるクラスタリング 10 回の平均正答率，棒グラフは，データ圧縮にかかる時間と，k-medoid によるクラスタリング 10 回の平均実行時間の和となっている．

DTWS 木による例数圧縮した結果，α の値によっては，圧縮無しのクラスタリングよりも正答率が良くなった．これは，似た形の例を圧縮しクラスの特徴をつかんだ平均時系列を作成したことによる．また，実行時間において，例数圧縮無しの場合と比べるとおよそ 15 から 60 倍速くなっている．縦方向に伸ばす平均時系列は，凸部分に特徴のあるクラスに関して正答率の向上が見られた．

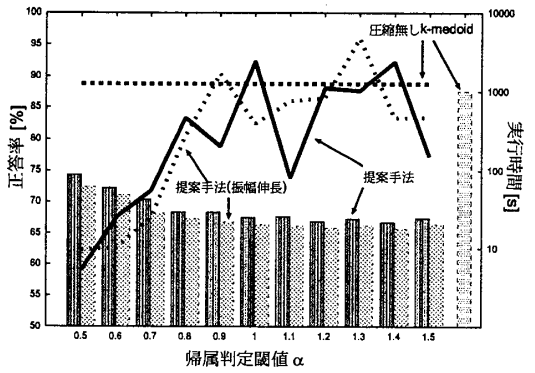

図 1: ”forget, innocent, lose” のクラスタリング結果

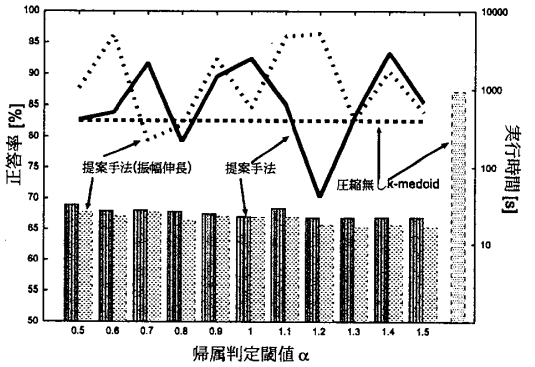

図 2: ”forget, later, Norway” のクラスタリング結果

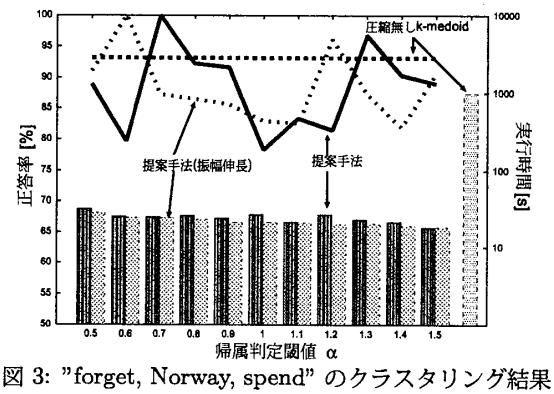

図 3: ”forget, Norway, spend” のクラスタリング結果

参考文献

[Zhang 96] T. Zhang, R. Ramakrishnan, and M. Livny; “BIRCH: An Efficient Data Clustering Method for Very Large Databases”, *Proceedings of ACM International Joint Conference on SIGMOD*, pp. 103-114, 1996.

[Keogh 00] E.J. Keogh and M.J. Pazzani; “Scaling up Dynamic Time Warping for Datamining Application”, *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 285-289, 2000.

[Sakoe 78] H. Sakoe and S. Chiba; “Dynamic Programming Algorithm Optimization for Spoken Word Recognition”, *IEEE Transaction on Acoustics, Speech, and Signal Processing*, Vol. ASSP-26, No. 1, pp. 43-49, 1978.

[Bay 99] S. Bay; “UCI Repository of KDD Database”, 1999, http://kdd.ics.uci.edu/.