I_023

ナイストロム法を用いた時系列データの高速類似検索法の検討

Fast similarity search of time series data using the Nyström method

澤村 康匡 † 林 朗 † 末松 伸朗 †
Sawamura Yasumasa Hayashi Akira Suematsu Nobuo

1 はじめに

1.1 時系列データの類似検索問題

近年，大量の時系列データからのデータマイニングが関心を集めており，そのための要素技術である時系列データの類似検索，類似部分列検索，分類，クラスタリング，セグメンテーションなどが盛んに研究されている．なかでも，時系列データの類似検索は直接的な利用だけでなく，分類，セグメンテーションなどの前処理として利用可能であり重要である．本研究ではベクトル値時系列データの類似検索に着目し，その高速化手法を検討する．問題設定は以下のとおりである．

1.2 問題設定

n 個の時系列データからなる集合（データベース (DB) 時系列）$\mathcal{X} = \{X_1, \ldots, X_n\}$ が与えられたとする．ここで，$X_i (1 \leq i \leq n)$ は特徴ベクトルから構成される長さ l_i の系列 $X_i = (\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_{l_i}^i)$ である．このとき，新たに与えられた時系列データ（クエリ時系列）$Q = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_l)$ の N_{br} 近傍すなわち，DTW 距離 $D(Q, X_i)$ を最小にするような DB 時系列 $X_i \in \mathcal{X}$ を N_{br} 個，効率的に求めよ．

1.3 提案手法

類似検索を行うためには各 DB 時系列とクエリ時系列間の反類似度の定義が必要となる．反類似度としてはユークリッド距離や動的時間伸縮（Dynamic Time Warping, DTW）距離などが用いられている [1]．DTW は2つの特徴ベクトル時系列に対し時間伸縮を許して可能な全ての対応を評価し，その中から類似度最大となる対応付けを見出すものであり，時間軸の固定されたユークリッド距離に比べて時間軸方向のずれに頑健であり，より直感的な類似度を反映していると考えられる．本研究では DTW 距離を反類似度として，時系列データを低次元ユークリッド空間に埋め込み，埋め込み空間内で多次元探索を行うアプローチを検討する．

多次元探索はユークリッド距離を用いる近傍探索では効率的な手法として広く知られているが時系列データを対象として DTW 距離を用いる近傍探索は DTW 距離の計算量（時系列長の2乗時間）が大きい，距離の公理の1つである三角不等式 $(D(X,Z) \leq D(X,Y) + D(Y,Z))$ を必ずしも満たさない，という問題点により，そのままの適用は困難である．

効率的な検索システムを実現するには，いかに少数の DTW 距離から精度よく埋め込みを行えるかが重要となる．本研究では，距離から求まるカーネル関数により定まる積分方程式を考え，積分方程式の数値解法であるナイストロム法 [5] に基づき，クエリ時系列と少数のサンプル時系列間の DTW 距離から時系列データを高い精度で補間埋め込みする手法を提案する．

埋め込み手法の候補として，多次元尺度法（MDS）[3]，およびラプラシアン固有マップ法（LE 法）[5] を検討する．

1.4 関連研究

Yi らは FastMap を用い，DTW 距離に基づき時系列データをユークリッド空間へ埋め込み，埋め込み空間内で多次元探索を行うことを提案した [4]．FastMap はヒューリスティックを用いて MDS を高速化した手法である．距離から求まるカーネル行列の固有ベクトル（主成分ベクトル）を近似するようなピボットとよぶ対のデータを選び，ピボット対を結ぶ軸により埋め込み空間をはり，軸への射影により埋め込み座標を求める．FastMap を用いることで，DTW 距離の計算を減らすことができるが，埋め込み次元数によりピボット数が決まるため，低次元空間への高い精度の埋め込みは困難である．

多次元探索が困難であるため，線形探索を前提として DTW 距離の下界を高速に計算する研究が行われている [1]．クエリ時系列からの DTW 距離が ϵ 以上のものを予め検索対象から除外するのが目的である．しかし，これらの研究はスカラ値時系列に限定され，ベクトル値時系列へは適用できない．

2 埋め込み手法

2.1 MDS

時系列間 DTW 距離を $\{d^2(X_i, X_j) \mid 1 \leq i, j \leq n\}$ とする．MDS[3] は，

$$\|\Phi(X_i) - \Phi(X_j)\|^2 = d^2(X_i, X_j) \quad (1 \leq i, j \leq n) \qquad (1)$$

を満たすような写像 $\Phi : \mathcal{X} \to \Re^n$ による X_i の埋め込み座標 $\boldsymbol{z}_i$ を間接的に求める手法である．

中心化した内積（カーネル）行列 $k(X_i, X_j)$ を，$\overline{\boldsymbol{z}} = \frac{1}{n}\sum_i \boldsymbol{z}_i$，$k(X_i, X_j) = \langle \boldsymbol{z}_i - \overline{\boldsymbol{z}}, \boldsymbol{z}_j - \overline{\boldsymbol{z}} \rangle$ とすれば，ユークリッド空間における内積と距離の関係より，下記の式 (2) が求まる [2]．

$$k(X_i, X_j) = -\frac{1}{2} d^2(X_i, X_j) + \frac{1}{2n}\sum_{l=1}^{n} d^2(X_i, X_l) + \frac{1}{2n}\sum_{l=1}^{n} d^2(X_j, X_l) - \frac{1}{2n^2}\sum_{l=1}^{n}\sum_{m=1}^{n} d^2(X_l, X_m) \qquad (2)$$

カーネル行列 $\boldsymbol{K}$: $\boldsymbol{K}_{ij} = k(X_i, X_j)$ の固有値解析を行い，$\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ のように分解する．ただし，固有値行列

† 広島市立大学大学院情報科学研究科 〒731-3194 広島市安佐南区大塚東 3-4-1
Email: sawamura@robotics.im.hiroshima-cu.ac.jp

$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \lambda_1 \geq \ldots \geq \lambda_n$，固有ベクトル行列 $\boldsymbol{U} = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n]$ である．

$\boldsymbol{K}$ が半正定行列である場合には，$\boldsymbol{Z} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^T$ とおくと，$\boldsymbol{K} = \boldsymbol{Z}^T\boldsymbol{Z}$ となる．ここで，$\boldsymbol{Z}$ の i 列は，$\boldsymbol{z}_i - \bar{\boldsymbol{z}}$ であると考え，重心位置に原点を移動したものを改めて $\boldsymbol{z}_i (1 \leq i \leq n)$ とする．$\boldsymbol{z}_i$ に対して，n 次元空間から p 次元空間への射影を考える．その射影による $\boldsymbol{z}_i$ の像 $\tilde{\boldsymbol{z}}_i$ は，カーネル行列 $\boldsymbol{K}$ の大きい方から p 個の固有値，固有ベクトルを用いて，以下のように表される．

$$\tilde{\boldsymbol{z}}_i = (\sqrt{\lambda_1}\boldsymbol{e}_1(i), \sqrt{\lambda_2}\boldsymbol{e}_2(i), \ldots, \sqrt{\lambda_p}\boldsymbol{e}_p(i))^T \quad i = 1, \ldots, n \tag{3}$$

ここに，$\boldsymbol{e}_k(i)$ は固有ベクトル $\boldsymbol{e}_k$ の i 番目の要素である．

しかし，DTW 距離は前述の通り距離の公理を満たさないため，(2) 式から定まる行列 $\boldsymbol{K}$ は必ずしも半正定値とは限らない．そこで，負の固有値，固有ベクトルを無視して，(3) 式によりユークリッド空間への埋め込みを行う．

2.2 ラプラシアン固有マップ法

時系列データ間の近傍関係を保存するようにユークリッド空間に埋め込む手法として，ラプラシアン固有マップ法（LE 法）[5] が提案されている．距離が三角不等式を満たさなくとも，ラプラシアン行列に負の固有値が出てこない点が特長がある．

まず，類似度行列 $\boldsymbol{W}$ を (4) 式により計算する．ここで，$N_k(X_i)$ は X_i の k 近傍であり，$t(>0)$ はハイパーパラメータである．

$$\boldsymbol{W}_{ij} = \begin{cases} 1 & i \neq j \bigwedge X_j \in N_k(X_i) \\ 0 & \text{上記以外のとき} \end{cases} \tag{4}$$

$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ で表されるラプラシアン行列 $\boldsymbol{L}$ を計算する．$\boldsymbol{D}$ は $\boldsymbol{D}_{ij} = \Sigma_{j=1}^{n}\boldsymbol{W}_{ij}$ なる対角行列である．なお，行列 $\boldsymbol{L}$ は半正定値である [2]．$\boldsymbol{L}$，$\boldsymbol{D}$ を用いて次の一般化固有値問題を解く．

$$\boldsymbol{L}\boldsymbol{e} = \lambda\boldsymbol{D}\boldsymbol{e} \tag{5}$$

得られた固有値の小さなものから p 個の対応する固有ベクトルを用いて，時系列 X_i の埋め込み座標 $\boldsymbol{z}_i$ を (6) 式のように表す．

$$\tilde{\boldsymbol{z}}_i = (\boldsymbol{e}_1(i), \boldsymbol{e}_2(i), \ldots, \boldsymbol{e}_p(i))^T \quad i = 1, \ldots, n \tag{6}$$

3 ナイストロム補間

ナイストロム法は，第二種フレッドホルム積分方程式の数値解を得るための手法である [5]．ナイストロム補間とはサンプル集合 $\hat{\mathcal{X}}$ に含まれない点が与えられたとき，(7) 式によって，その点における固有関数の値 $f_k(\boldsymbol{x})$ を改めて固有値問題を解き直すことなく近似計算する方法である．

$$f_k(\boldsymbol{x}) = \frac{1}{m\lambda_k'}\sum_{i=1}^{m}\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{x}_i)f_k(\boldsymbol{x}_i). \tag{7}$$

まずサンプル点 $X_i \in \hat{\mathcal{X}}$ に対して定義されるカーネル関数 ((2) 式) の定義域を次式のように任意の二点 X_I, X_J に対して拡張する．

$$k(I,J) = -\frac{1}{2}d^2(I,J) + \frac{1}{2m}\sum_{l=1}^{m}d^2(I,l) + \frac{1}{2m}\sum_{l=1}^{m}d^2(J,l) - \frac{1}{2m^2}\sum_{l=1}^{m}\sum_{l'=1}^{m}d^2(l,l') \tag{8}$$

ナイストロム拡張の MDS への適用は，予め固有値解析を行い，サンプル点での固有関数の値 $f_k(X_i) \quad i = 1, \ldots, m, \quad k = 1, \ldots, p$ を求めた後，以下のように行う．

1. $X_q \notin \hat{\mathcal{X}}$ が与えられる．
2. 式 (8) により，$k(q,i), \quad i = 1, \ldots, m$ を求める．
3. 式 (7) により，$f_k(X_q), \quad k = 1, \ldots, p$ を求める．
4. $\boldsymbol{z}_k(X_q) = \sqrt{\lambda_k'}f_k(X_q), \quad k = 1, \ldots, p$ を埋め込み座標とする．

LE 法の場合，一般化固有値問題が (7) 式へ直接対応しないため上記の式をそのまま適用できない．ここでは Benjio らによるナイストロム拡張のスペクトラルクラスタリング（SC 法）への適用 [5] に基づいた導出を行う．SC 法では，次式で定義されるカーネル行列 $\boldsymbol{K}$ の固有値問題を解く．

$$\boldsymbol{K} = \boldsymbol{D}^{\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{\frac{1}{2}} \tag{9}$$

ここに，$\boldsymbol{W}$ と $\boldsymbol{D}$ は LE 法に対して定義されたものと同じである．ナイストロム拡張の SC 法への適用は，以下のカーネル関数に基づいて行う．

$$k(I,J) = \frac{\boldsymbol{W}(I,J)}{\sqrt{\boldsymbol{D}(I,I)}\sqrt{\boldsymbol{D}(J,J)}} \tag{10}$$

$$\boldsymbol{D}(I,I) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}(X_I, X_i)$$

式 (5) の固有値，固有ベクトルを λ_k^{LE}，$\boldsymbol{e}_k^{LE}$ とすれば，SC 法の一般化固有値 $\lambda_k^{SC} = 1 - \lambda_k^{LE}$，固有ベクトル $\boldsymbol{e}_k^{LE} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{e}^{LE}$ となることが容易に示せる．従って，ナイストロム拡張の適用は予めサンプル点に対する (9) 式の固有値解析を行い，MDS と同様に固有関数の値 $f_k(X_i)$ を求めて，以下のように行う．

1. $X_q \notin \hat{\mathcal{X}}$ が与えられる．
2. 式 (10) により，$k(q,i), \quad i = 1, \ldots, m$ を求める．
3. 式 (7) により，$f_k(X_q), \quad k = 1, \ldots, p$ を求める．
4. $\boldsymbol{e}_k^{SC} = (f_k(X_1), \ldots, f_k(X_m), f_(X))^T$ として，LE 法に対応する解 $\boldsymbol{e}_k^{LE}, \quad k = 1, \ldots, p$ を得る．
5. $\tilde{\boldsymbol{z}}_k(X_q) = \boldsymbol{e}_k^{LE}(m+1), \quad k = 1, \ldots, p$ を埋め込み座標とする．

4 提案手法

提案手法の類似検索手順は前処理である DB 時系列の作成と，クエリ時系列の検索の 2 つのフェースで構成される．

■前処理フェーズ

1. DB 時系列の集合 $\mathcal{X} = \{X_1, \ldots, X_n\}$ が与えられる．
2. $\mathcal{X}$ から m 個 ($m \ll n$) のサンプルを選択し，添え字を付け直して $\hat{\mathcal{X}} = \{X_1, \ldots, X_n\}$ とする．
3. サンプル間の DTW 距離 $d_{tw}(X_i, X_j) \quad (1 \leq i, j \leq m)$ を計算する．
4. すべてのサンプル $X_i \in \hat{\mathcal{X}}$ を埋め込み手法でユークリッド空間に埋め込み $\{\tilde{\boldsymbol{z}}_i | 1 \leq i \leq m\}$ を求める．
5. サンプルを除く $(n-m)$ 個の DB 時系列 $X_i \quad (m+1 \leq i \leq n)$ をナイストロム法により補間埋め込みをし，$\{\tilde{\boldsymbol{z}}_i | m+1 \leq i \leq n\}$ を求める．

6. $\{\tilde{z}_i|1 \leq i \leq n\}$ について多次元インデックスを構築する．

■検索フェーズ

1. クエリ時系列 $Q \notin \mathcal{X}$ を取得する．
2. クエリ時系列 Q と m 個のサンプル間の DTW 距離 $d_{tw}(Q, X_i)$ $(1 \leq i \leq m)$ を計算する．
3. クエリ時系列 Q をナイストロム法で補間埋め込みし，$\tilde{z}_Q$ を求める．
4. $\tilde{z}_Q$ をキーとして，$\{\tilde{z}_i|1 \leq i \leq n\}$ の多次元探索を行ない，近傍 $N_{br}(Q) = \{X_i|\tilde{z}_i \in N_{br}(\tilde{z}_Q), 1 \leq i \leq n\}$ を求める．
5. 求まった近傍内の DB 時系列 $X_i \in N_{br}(Q)$ とクエリ時系列 Q との DTW 距離を計算し，真に近しいものを検索結果として返す．

5 補間埋め込みを用いた際の計算量

まず時系列長さ (平均 l) とし，DB 内の時系列データの数 n として p 次元空間に埋め込む場合の計算量を考える．DTW 距離の計算に $O(nl^2)$，データの埋め込みに $O(n^3p)$ の計算量がかかる．補間埋め込みを用いた際の計算量は，サンプル数 m とすれば，DTW 距離の計算に $O(ml^2)$ サンプルの埋め込みに $O(m^3p)$ だけかかる．補間埋め込み自体の計算量は $O(mnp)$ で計算することができる．従って，ナイストロム補間埋め込みを適応した場合の埋め込み計算量は，$O(ml^2) + O(m^3p) + O(mnp)$ となる．ナイストロム補間を用いることで計算量の項数は増えているが，$m < n$ である為，ナイストロム補間を用いた方が計算量が少なく，n に比べて m が小さいほど高速化につながる．

6 実験

埋め込み手法（MDS，LE 法）の比較評価，およびサンプル数の埋め込み精度への影響の評価を目的として，スカラ値の人工時系列データと，ベクトル値の実時系列データの 2 種類のデータに対して実験を行った．

類似順位上位 N_{true} 位までの時系列データを検索するタスクに取り組み，各埋め込み手法について，異なるサンプル数の下での再現率適合率曲線を求める．

再現率および適合率は情報検索の分野でよく用いられる指標である．埋め込み空間内で第 N_{search} 近傍までを検索結果として，その中に含まれる正解 (DTW 距離による第 N_{true} 近傍までの時系列) の数 N_{find} を N_{search} の値を変えながら調べ，結果を以下の再現率 (recall)R，適合率 (precision)P によって表した．

$$R = \frac{N_{find}}{N_{true}}, \qquad P = \frac{N_{find}}{N_{search}} \tag{11}$$

■実験 1　CBF(Cylinder-Bell-Funnel) は時系列データマイニングの分野でよく使われている人工データである [1]．CBF は乱数を用いて合成されるスカラ値時系列であり，3 つのクラスがある．今回はすべての時系列長 $l = 32$ として，以下の合成式によりデータを作成した．

$$\begin{aligned} \boldsymbol{c}(l) &= (8+z)R[a,b] + \boldsymbol{e}(l) \\ \boldsymbol{b}(l) &= (8+z)R[a,b]\frac{l-a}{b-a} + \boldsymbol{e}(l) \\ \boldsymbol{f}(l) &= (8+z)R[a,b]\frac{b-l}{b-a} + \boldsymbol{e}(l) \end{aligned} \tag{12}$$

$$R[a,b] = \begin{cases} 0 & l < a \quad or \quad l > b \\ 1 & a \leq l \leq b \end{cases} \tag{13}$$

ただし，z，$\boldsymbol{e}(l)$ は平均 0，分散 1 の正規分布からサンプリングし，a（整数）は区間 $[4,8]$，$(b-a)$（整数）は区間 $[8,24]$ の範囲から一様にサンプリングした．

得られた 3 つのクラス c，b，f のデータをそれぞれ 4000 個，3000 個，3000 個用いて合計 10000 個を DB 時系列とし，別に作成したクラス c のデータ 100 個をクエリ時系列として MDS，LE 法を用いて 50 次元に埋め込み，10 近傍探索を行った．補間埋め込みに用いるサンプル数は DB 時系列の 50%，25%，10%，5% とした．

図 1 に再現率適合率曲線を，各埋め込みデータに対し再現率が約 90% 時の適合率の値を表 1 にそれぞれ示す．

表 1 より，再現率 $R = 90\%$，すなわち 10 最近傍のうち 9 個を見つけるためには，DB 時系列全体の 50% のサンプル数，5000 個でナイストロム補間した LE 法で，適合率 $R = 5.45\%$，約 170 近傍探す必要がある．MDS ではサンプル数 5000 のとき，適合率 $R = 2.07\%$ で約 440 近傍まで探す必要があるが，サンプル数が少なくなっても適合率の値はあまり変わらない．5% のサンプル数でも約 480 近傍まで探索すればよい．

表 1　実験 1 における再現率 R(%)，適合率 P(%) の比較

埋め込み手法	Sample	R	P
LE 法	500	90.1	0.979
	1000	90.2	2.51
	2500	90.1	4.90
	5000	90.4	5.45
MDS	500	90.5	1.97
	1000	90.7	1.97
	2500	91	2.07
	5000	91.2	2.07

■実験 2　実ベクトル値時系列データとしてオーストラリア手話言語，ASL を用いて近傍探索を行った．5 人の被験者から得た（1 単語あたり平均 73 個の）手話データであり，それぞれ 9 次元特徴ベクトル〜構成され，データ毎に時系列長が異なる．実験では"change"，"deaf"，"her"，"innocent"など類似単語を持つ 43 単語を選び，DB 時系列とした．クエリ時系列として"lose"，"love"の 2 単語を使用した．

データ数 3000，クエリ数 100 として，MDS と LE 法により 30 次元に埋め込み，10 近傍探索を行った．採用したサンプル数は DB 時系列の 100%，50%，10%，5% である．結果を図 2，表 2 に示す．

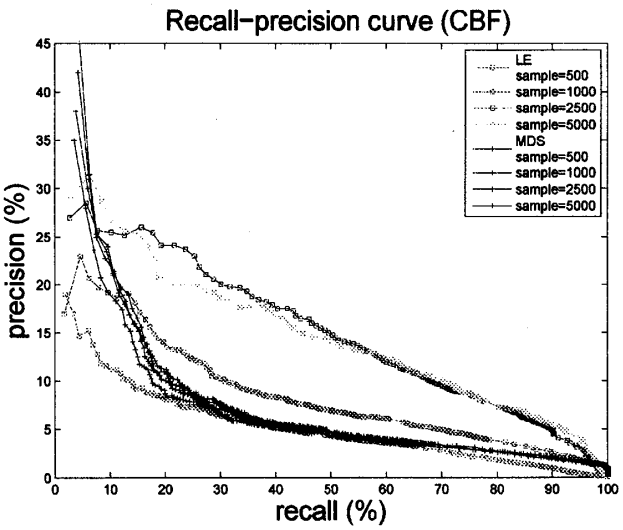

図1　CBF 10近傍探索における再現率・適合率

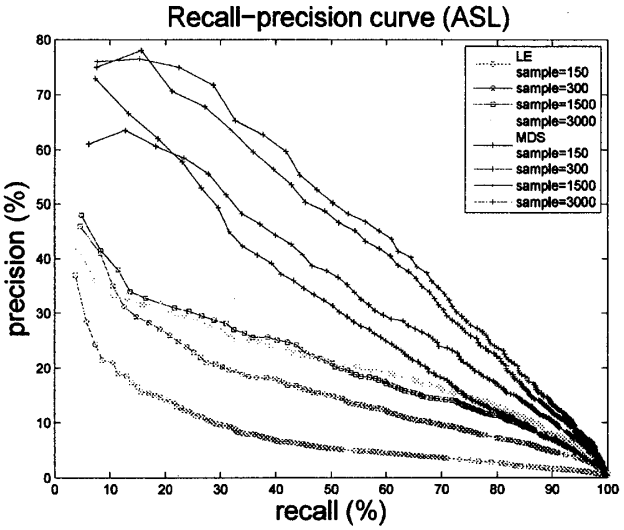

図2　ASL 10近傍探索における再現率・適合率

表2より，10最近傍のうち9個を見つけるためには補間なしのLE法で適合率 $R = 7.82\%$，約115近傍をMDSでは適合率 $R = 13.3\%$，約70近傍までそれぞれ探す必要がある．ナイストロム補間の結果はLE法，MDSとも50%のサンプル数では精度はほとんど下がらず，サンプル数5%についてみると，MDSは約130近傍まで探す必要があるが比例的な精度の減少をしていないことが確認できる．LE法では約560近傍まで探索する必要があり，MDSに比較して埋め込み精度の低下が激しいのが確認できる．

表2　実験2における再現率R(%)，適合率P(%)の比較

埋め込み手法	Sample	R	P
LE法	150	90.5	1.62
	300	90.2	4.75
	1500	90.2	7.52
	3000	90.7	7.82
MDS	150	90.1	6.93
	300	90.2	10.1
	1500	90.1	11.9
	3000	90.1	13.3

7　考察

今回用いた2つの埋め込み手法について考察する．実験1についてみると，LE法は短い距離に重きをおいて近傍関係を保存するように埋め込むため，サンプル数が多いときではMDSより高い精度が得られているが，サンプル数が少なくなるにつれ精度が悪化している．これは近傍関係を保存するのに十分なサンプル数が得られなくなったためと考えられる．

実験2ではMDSでは比較的少数のサンプルからも精度よく埋め込めているが，これは選んだ少数のサンプルから全データを用いて得られる埋め込み空間を再現できたためと考えられる．

8　まとめ

時系列データの高速類似検索手法として，低次元ユークリッド空間に埋め込みを行い，埋め込み空間内で多次元探索を行う方法を提案した．非類似度としてDTW距離を用い，低次元埋め込み手法としてMDS，LE法を検討し，計算負荷を軽減するためにナイストロム補間を組み合わせて，実験により埋め込み結果の比較評価を行った．実験の結果，実データであるASLに関しては少数のサンプルデータでも，あまり精度を落とさず類似検索が行えることが確認できた．

また，今回行った実験ではいずれも30次元をこえる高次元空間へ埋め込みを行った．多次元探索が効率的に行えるのは10数次元であるといわれており，このような高次元空間内のデータに対して効率的な探索方法を考えることが今後の課題として挙げられる．

参考文献

[1] E. Keogh. Exact indexing of dynamic time warping, 2002.

[2] 水原 悠子. DTW距離を用いた時系列データのベクトル空間への埋込, 信学論, pp. 241–249, 2005.

[3] W. S. Torgerson. Theory and methods of scaling, J. Wiley, 1958.

[4] B. Yi et al. Efficient Retrieval of Similar Time Sequences Under Time Warping, ICDE, pp.201–208, 1998.

[5] Y. Bengio et al. Learning eigenfunctions links spectral embedding and kernel PCA, Neural Computation, pp.2197–2219, 2004.