Vol.2012-MPS-91 No.27
Vol.2012-BIO-32 No.27
2012/12/7

Local Shapeletを用いた時系列分類に最適な距離尺度の選択

辻本 貴昭[1,a)]　上原 邦昭[1,b)]

概要：時系列データに対する距離尺度は多く提案されているが，1 種類のみの距離尺度ではデータセットの類似性を正確に評価できるわけではない．本稿では，Local Shapelet を用いて，各データセットごとに性能が良いと考えられる距離尺度を選択する枠組みを提案する．そして，各データセットごとに適切な距離尺度を選択すれば，単一の距離尺度を用いてデータの類似性を評価するよりも，様々なデータセットに対して性能が良くなることを示す．

On the Selection of Appropriate Distance Measure for Timeseries Classification using Local Shapelet

Takaaki Tsujimoto[1,a)]　Kuniaki Uehara[1,b)]

Abstract: There exists many distance measures for time series data. However, none of them can estimate similarity correctly on all datasets. This paper describes a framework for selecting an appropriate distance measure for each dataset using Local Shapelet. Experimental results with the selected distance measure show better performance than with other distance measures.

1. はじめに

従来より，時系列データの分類はデータ解析の観点から多くの注目を集めている．例えば，株価の動きからの株価の予測や，音声認識，画像認識などに用いられている．これらの分野では，時系列データどうしを比較し，その類似性を評価することが頻繁に行われる．このうち，もっとも古くから利用されている手法の一つとして EUC (ユークリッド距離) がある．EUC は，長さ n の時系列に対して n 回距離を計算し，足しあわせて比較する．しかし，時系列の各点間の距離は独立して比較されるため，長さの異なる時系列を比較することは難しく，また，ノイズに対応できないという問題があった．そのため，時系列データどうしの距離を測るために，数多くの距離関数が提案されている．

これまでに提案されてきた各距離尺度は，特定の分野のデータに対してのみ優れている場合が多く，データセットによって性能が良い距離尺度が異なることが知られている．このため，H. Ding[1] らは，様々なデータセットに対して多くの距離尺度を用いて性能比較した．この実験により，単一の距離尺度が全てのデータセットに対して性能が良いということはなく，多くの距離尺度は，40 年前に提案された DTW (Dynamic Time Warping)[2] とほとんど性能が変わらないということがわかった．さらに，DTW 以外の距離尺度が無意味なのではなく，特定のデータセットに関しては，性能が良い固有の距離尺度が存在することも確認された．以上のことから，全てのデータに対して同じ距離尺度で距離を計算するのではなく，データセットごとに最適な距離尺度を用いる方が良いと考えられる．

データセットにより性能が良い距離尺度が異なるという事は，データセットの特徴によって性能が良い距離尺度を決定すればよいということになる．このようなデータセットの特徴を表現するために Local Shapelet[3] を用いる．そして，Locah Shapelet を用いてデータセットごとに最適

[1] 神戸大学大学院 システム情報学研究科
Graduate School of System Informatics, Kobe University
[a)] 2gmon@ai.cs.scitec.kobe-u.ac.jp
[b)] uehara@kobe-u.ac.jp

と考えられる距離尺度を選択する方法を提案する．Local Shapelet とは，同じクラスに共通する部分時系列である．時系列データがこの部分時系列と似たデータを含むかどうかを EUC を用いて判定する．そして，もし部分時系列を含んでいると判定されるならば，時系列データは Local Shapelet の属するクラスに分類される．

一方，Web 検索などに用いられるランキング学習 [4] では，ドキュメントにクエリが何個含まれるか，クエリに関係したページからのリンク数が何個あるかなど，複数のパラメータの重みを学習し，ドキュメントに対するクエリの適合度という計算結果から，ドキュメントの順位を求めている．本稿では，ランキング学習の計算結果からドキュメントの適合順位を決定するというアイデアを借りて，分類エラー率という結果から，データセットに対して性能が良い距離尺度の優先順位を決定することを提案する．具体的には，各距離尺度を用いてそれぞれのデータを分類した際のエラー率という一種の計算結果がわかっている．エラー率をランキング学習における適合度とみなせば，エラー率からそれぞれの距離尺度が得意なデータセット，苦手なデータセットを判断することができる．したがって，分類対象となっているデータセットに対して一番性能が良いと考えられる距離尺度を選択すれば，単一の距離尺度を用いるよりも，様々なデータセットに対して性能良く分類できることになる．

2. 関連研究

2.1 距離尺度

時系列データの類似性を評価するために用いられる距離尺度の中で，最も単純なものは EUC である．EUC はデータの時間方向への伸縮に対応できないため，DTW が提案された．DTW はお互いのデータ間の距離が最小になるように時間軸を伸縮させる手法である．この手法により，長さの異なるデータどうしも比較することが可能になる．

DTW は各時間での値の差が最小になるように時間軸を伸縮させるので，データによっては一部を異常に引き伸ばしてしまい，データの類似性が正しく評価できないことがある．このようなデータの類似性を正しく評価するために，各時間ごとのデータの変化量に着目した DDTW (Derivative Dynamic Time Warping)[5] が提案された．一方，移動軌跡データなど，異常値を多く含むデータは，既存の類似尺度では類似性を正確に評価できないので，LCSS（distance based on Longest Common Subsequence)[6]，EDR (Edit Distance on Real sequence)[7] などの編集距離に基づいた距離尺度が提案された．編集距離に基づく距離尺度は，各時間での値の差が閾値以上の場合のみ，距離が 1 だけ離れていると判断する．したがって，僅かな差は距離として認識されず，逆に異常値などの大きすぎる差も距離 1 としか認識しない．これによって異常値を多く含むようなデータの類似性も正確に評価することが可能となる．

このように，時系列データの距離尺度として様々な手法が提案されている．そして，それらの距離尺度は特定のデータセットのみで優れており，そのデータセットで既存の距離尺度に対する優位性を主張している．しかし，ある距離尺度は別の距離尺度よりも優れているという主張や，その逆に，両方共あまり変わらないなどといった矛盾した論文も多く存在している [8]．そのため，様々な分野のデータに対して各距離尺度の優位性を確認するために，H. Ding らは実験により多くの距離尺度を比較した．この実験により，

(1) 訓練データが多い場合, EUC と, DTW, LCSS, EDR, ERP (Edit Distance with Real Penalty)[9] などのタイムワーピングを用いた距離尺度の性能には差が無い．
(2) LCSS，EDR，ERP などの編集距離に基づいた距離尺度の性能は，40 年前に提案された DTW の性能とほとんど違いがないが，EDR のみ少し性能が良い可能性がある．
(3) TQuEST (similarity search based on Threshold Queries)[10]，SpADe (Spatial Assembling Distance)[11] などの新しい距離尺度は，一般的に，タイムワーピングを用いた距離尺度よりも劣っている．

といったことがわかっている．したがって，全てのデータに対して一番良いと言える距離尺度は存在せず，正確に類似性を評価するためには，データセットごとに性能が良い距離尺度を選択しなければならない．以上のことから，本稿では，対象とするデータごとに，そのデータセットに対して性能が良い距離尺度を選択する方法を提案する．

2.2 Local Shapelet

時系列データの特徴は，平均や分散などの統計的な数値を用いて表現することもできるが，統計的な特徴では実際のデータを想像することが難しい．そこで，時系列データの特徴を表現する方法として Local Shapelet を用いる．Local Shapelet は，元データの部分時系列を直接的に用いるので，データの特徴を視覚的に判断しやすいというメリットがある．

Local Shapelet は，部分時系列，距離の閾値，ターゲットクラスの 3 つからなる．ある時系列データの部分時系列と Local Shapelet の距離が閾値以下の場合，その時系列データのクラスが Local Shapelet のターゲットクラスである確率は x 以上であるという．この x は任意に設定することができ，Local Shapelet の部分時系列と，全ての訓練データから得られる部分時系列の距離を比較すれば，x に基づいて閾値が統計的に決定される．そして，数多くの候補の中から，ターゲットクラスに一番多く含まれていると考えられる Local Shapelet を選択し，そのターゲットクラスの特徴とする．

3. Local Shapelet を用いた時系列分類に最適な距離尺度の選択

3.1 代表距離尺度の決定

まず，簡単のために，様々な距離尺度の中から代表となる距離尺度を決定する方法について述べる．距離尺度を選択する際には，各データセットに対して，類似性を最も正確に評価できるような距離尺度を組み合わせとして選ばなければならない．つまり，得意なデータセットができるだけ重ならない距離尺度を選ばなければならない．なぜなら，同じデータセットに対して得意な距離尺度ばかりを選択してしまうと，他の苦手なデータセットの類似性は正確に評価できなくなるからである．

それぞれの距離尺度が得意なデータセットの情報として，H. Ding らの実験結果の分類エラー率がある．この結果の中から，本稿で用いた 20 のデータセットを抜き出したものを表 1 に示す．この表で比較されている距離尺度のうち，エラー率が低いのは得意なデータセット，エラー率が高いのは苦手なデータセットだと考えられる．そこで，ドキュメントのクエリに対する適合度という計算結果から順位を決定するというランキング学習のアイデアを借りて，データセットに対して性能が良い距離尺度の優先順位を決定する方法を提案する．

ランキング学習とは，Web 検索などに用いられる手法で，ドキュメント中に含まれるクエリの数や，クエリに関係したページからのリンク数などの，複数のパラメータの重みを学習し，クエリとドキュメントの適合度を求める手法である．重みの学習の際には，クリックログなどを利用して，クエリに良く適合するドキュメントの適合度が高くなるようにする．各クエリの適合度が同じようなドキュメントどうしは似ていると考えられるので，それぞれの距離尺度のデータセットに対する分類エラー率という計算結果から，得意なデータセットが同じ距離尺度を求めることを考える．具体的には，複数の異なるデータセットに対して，似た分類エラー率を持つ距離尺度どうしは，得意なデータセット，あるいは苦手なデータセットが似た距離尺度とみなすことができるので，それらの距離尺度をまとめることを考える．表 1 の列を，ドキュメント，行をクエリとみなすと，エラー率は学習された適合度とみなすことができる．そして，表 1 は 20 次元の 12 個の事例とみなすことができる．この 12 個の事例をクラスタリングすれば，得意なデータセット，苦手なデータセットの似た距離尺度をクラスタとしてまとめることができる．そして，それぞれのクラスタから一番平均性能が良い距離尺度を 1 つずつ選択すれば，得意なデータセットができるだけ重ならない距離尺度をいくつか選択することができる．

3.2 Local Shapelet の抽出

次に，Local Shapelet の抽出について述べる．Local Shapelet は，本来，Early Classification に用いられる特徴である．Early Classification とは，精度を一定以上に保ったまま，できるだけ初期に分類をすることを目的としている．したがって，時系列データの前半に現れる，クラスに特徴的な部分時系列を Local Shapelet として，データの前半部のみを用いた分類が行われる．本稿では，各手法が得意な Local Shapelet を抽出し，データセットがその Local Shapelet を含んでいるかどうかで，適切な距離尺度を選択する．

まず, 各データセットに含まれる全ての訓練データに, そのデータセットに対して一番性能が良い代表距離尺度をラベルとして付与する．そして, 全ての訓練データから, 任意の長さの部分時系列を全て抽出して，Local Shapelet の候補とする．これらの候補のターゲットクラスは, 先ほど抽出元のデータに付与されたラベルとなる．Local Shapelet を構成する部分時系列 s と時系列データ t の距離 $BMD(s,t)$ (Best Match Distance) は以下の式で定義される．

$$BMD(s,t) = \min\{EUC(s,s'), s' \subseteq t, len(s) = len(s')\} \quad (1)$$

式 (1) は，s と，s と同じ長さの t の部分時系列 s' の距離のうち，最小の値を示している．従って，$BMD(s,t) = 0$ のとき，時系列データ t は Local Shapelet のターゲットクラスと分類できる．しかしながら，ターゲットクラスの全てのデータとの BMD が 0 になることはなく，多少の誤差が存在する．誤差を許容しなければ正確に分類できないので，許容する誤差として，BMD の閾値を決める必要がある．パラメータ x は，Local Shapelet を構成する部分時系列 s と，時系列データ t の距離 $BMD(s,t)$ が，この Local Shapelet の閾値以上離れているときに, t が Local Shapelet のターゲットクラスではない確率を表している．パラメータ x に基づいて Local Shapelet の閾値を決定する方法には，カーネル密度推定 [12] を用いる方法とチェビシェフの不等式 [13] を用いる方法の 2 つがある．カーネル密度推定を用いた方法の方が少し性能が良いと言われているが，計算量はカーネル密度推定を用いた場合 $O(n^2)$，チェビシェフの不等式を用いた場合 $O(n)$ である．また，性能の差も僅かなので，本稿では，チェビシェフの不等式を用いる方法を採用している．

チェビシェフの不等式とは，ある確率変数が平均 μ，分散 σ^2 の時，$P(|X-\mu| \geq k\sigma) \leq 1/k^2$ が成立するというものである．つまり，確率変数の実現値が，平均 μ から $k\sigma$ 以上離れる確率が $1/k^2$ 以下であるということである．ターゲットクラスの時系列データとの BMD は小さく，ターゲットクラス以外の時系列データとの BMD は大きいと考えられるので，ある時系列データとの BMD が，ターゲットクラス以外の時系列データとの BMD の平均から離れ

表 1 様々な距離尺度を用いた 1-NN による分類のエラー率
Table 1 Error rates of different distance measures using 1-NN.

Data Set	EUC	L1	Linf	DTW	wDTW	EDR	ERP	LCSS	wLCSS	FTSE	SpADe	DDTW
50words	0.407	0.379	0.555	0.375	0.291	0.271	0.341	0.298	0.279	0.500	0.341	0.291
Adiac	0.464	0.495	0.428	0.465	0.446	0.457	0.436	0.434	0.418	0.961	0.438	0.380
Beef	0.400	0.550	0.583	0.433	0.583	0.400	0.567	0.402	0.517	0.617	0.500	0.585
CBF	0.087	0.041	0.534	0.003	0.006	0.013	0.000	0.017	0.015	0.040	0.044	0.377
Coffee	0.193	0.246	0.087	0.191	0.252	0.160	0.213	0.213	0.237	0.446	0.185	0.177
ECG200	0.162	0.182	0.175	0.221	0.153	0.211	0.213	0.171	0.126	0.539	0.256	0.167
FaceFour	0.149	0.144	0.421	0.064	0.164	0.045	0.042	0.144	0.046	0.368	0.250	0.048
FaceAll	0.225	0.192	0.401	0.060	0.079	0.050	0.028	0.046	0.046	0.127	0.315	0.125
fish	0.319	0.293	0.314	0.329	0.261	0.107	0.216	0.067	0.160	0.857	0.150	0.090
GunPoint	0.146	0.093	0.186	0.140	0.055	0.079	0.161	0.098	0.065	0.346	0.007	0.006
Lighting2	0.341	0.251	0.389	0.204	0.320	0.088	0.190	0.199	0.108	0.288	0.272	0.273
Lighting7	0.378	0.286	0.566	0.252	0.202	0.093	0.287	0.282	0.116	0.545	0.557	0.431
OliveOil	0.150	0.236	0.167	0.100	0.118	0.062	0.132	0.135	0.055	0.717	0.207	0.262
OSULeaf	0.448	0.488	0.520	0.401	0.424	0.115	0.365	0.359	0.281	0.405	0.212	0.082
Swed.Leaf	0.295	0.286	0.357	0.256	0.221	0.145	0.164	0.147	0.148	0.667	0.254	0.088
syn.con.	0.143	0.146	0.227	0.019	0.014	0.118	0.035	0.060	0.075	0.146	0.150	0.430
Trace	0.368	0.279	0.445	0.016	0.075	0.150	0.084	0.118	0.142	0.273	0.000	0.000
TwoPat.	0.095	0.039	0.797	0.000	0.000	0.001	0.010	0.000	0.000	0.003	0.052	0.000
wafer	0.005	0.004	0.021	0.015	0.005	0.002	0.006	0.004	0.004	0.106	0.018	0.010
yoga	0.160	0.161	0.181	0.151	0.151	0.112	0.133	0.109	0.134	0.430	0.130	0.078
average	0.247	0.240	0.368	0.185	0.191	0.134	0.181	0.165	0.149	0.419	0.217	0.195

るほど，その時系列データがターゲットクラス以外である確率は小さくなると考えられる．

まず，Local Shapelet と全ての訓練データとの BMD を計算する．そして，チェビシェフの不等式を適用する集合を得るために，計算された BMD の中から，ターゲットクラス以外の訓練データとの BMD のみを選択する．選択された BMD を確率変数の実現値の集合とみなし，チェビシェフの不等式に基づいて，この集合の平均から，標準偏差の $k = \sqrt{1/(1-x)}$ 倍離れた値を Local Shapelet の閾値とする．ただし，この値が負の場合は閾値を 0 とする．

このようにして求められる Local Shapelet の数は非常に多くなる．これらの中には，BMD が閾値以下になる時系列データがほとんど無い Local Shapelet や，逆に，ターゲットクラス以外の時系列データとの BMD のほとんどが閾値以下になる Local Shapelet も含まれている．このような Local Shapelet は，クラスの特徴を表すことができないので，除去する必要がある．

除去するために，それぞれのターゲットクラスにつき，F 値が上位 y 個の Local Shapelet のみを残している．残された Local Shapelet は，ターゲットクラスのデータに共通する部分時系列である．つまり，各距離尺度の得意なデータセットに共通する部分時系列とみなすことができる．

3.3 データセットに最適な距離尺度の選択

最後に，Local Shapelet を用いて，各データセットに最適な距離尺度を選択する方法について述べる．データセットには，異なるターゲットクラスに属する Local Shapelet が複数混在している．そのうち，一番多く含まれる Local Shapelet のターゲットクラスを調べるために，まず，対象のデータセットに含まれるデータそれぞれに対して，全ての Local Shapelet との BMD を計算する．つぎに，データとの BMD が閾値以下となる Local Shapelet の数を用いて投票を行い，それぞれのデータの類似性を一番正確に評価できる距離尺度を決定する．そして，データセット全体で一番多く選択された距離尺度を，このデータセットの類似性を最も正しく評価できる距離尺度と決定する．同数の投票があった場合は，表 1 の平均性能が高い方の距離尺度を選択する．

例えば，DDTW，EDR，DTW それぞれに対して 3 個の Local Shapelet を抽出し，テストデータに対して BMD が表 2 のようになるとする．列は Local Shapelet を表しており，Local Shapelet とデータの BMD が閾値以下なら 1，それ以外は 0 としている．この例を用いて最適な距離尺度の選択法の例を示す．

まず，それぞれのデータに対して，1 ならばターゲットクラスに投票を行い，0 ならば投票しないとする．そのうち最も多く投票されたものをそのデータのターゲットクラスと決定する．たとえば，データ 1，2，4，6 は，ターゲットクラスが DDTW である数が一番多いので，DDTW の性能が良いデータと考えられる．同様に，データ 5 は DTW の性能が良いデータと考えられる．データ 3 については，EDR と DTW の 1 の数が等しいが，表 1 によると，EDR の方が DTW よりも性能が良いので，EDR の性能が良いデータと考えられる．最終的に，6 個のデータのうち，DDTW が性能が良いと考えられるデータの数が一番多くなるので，このデータセットに対しては DDTW が選択される．

4. 実験

以下の実験では，UCR Time Series Homepage[14] のデータセットのうち，20 種類を用いた．訓練データとして配布されているデータから Local Shapelet を抽出し，それらを用いて，テストデータに適切な距離尺度を選択した．

4.1 代表距離尺度の選択

本稿では，階層型クラスタリング手法によって，H. Ding らの実験結果をクラスタリングして代表距離尺度を選択した．表 1 の行を特徴，列を事例とみなして階層型クラスタリングを行った結果を図 1 に示す．結果を見ると，4 つのクラスタが構築されていることが分かる．しかし，Linf,

表 2 テストデータと Local Shapelet の比較結果
Table 2 Comparison of test data and Local Shapelet.

	DDTW	DDTW	DDTW	EDR	EDR	EDR	DTW	DTW	DTW
データ 1	1	1	1	0	0	1	0	1	0
データ 2	1	0	1	0	1	0	0	0	0
データ 3	0	0	0	0	1	1	1	1	0
データ 4	1	1	1	1	0	0	0	0	1
データ 5	0	0	0	0	0	1	1	1	1
データ 6	0	1	1	1	0	0	0	1	0

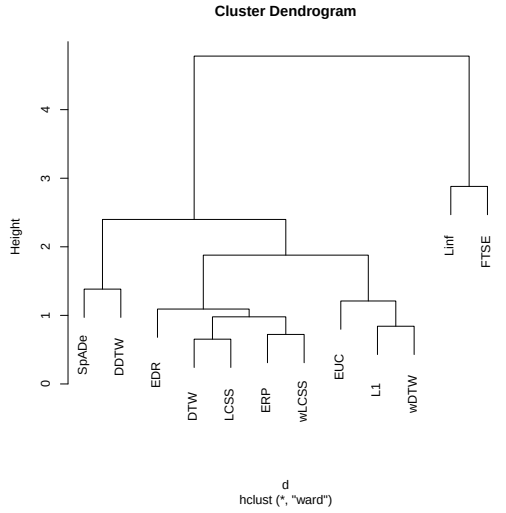

図 1 各距離尺度のエラー率のクラスタリング結果
Fig. 1 Clustering of error rates of each distance measure.

FTSE はほぼ全てのデータに対して良い性能が示されていないので，代表距離尺度から除外している．残りの 3 つのクラスタから，それぞれ平均エラー率が最も低い距離尺度を選択すると，DDTW，EDR，wDTW が選択される．

SpADe と DDTW は 1 つのクラスタを構築している．両距離尺度は，時間ごとの値の変化量に着目するなど，データの形状を重視しているとみなすことができる．他に，EDR，DTW，LCSS，ERP，wLCSS が 1 つのクラスタを構築している．DTW を除いた各距離尺度は，編集距離に基づいて提案された距離尺度である．そして，もう 1 つのクラスタを構築している，EUC，L1，wDTW は，データ間の値の差を重視している距離尺度と考えられる．この結果より，各時間での値の差を重要視する距離尺度，データの傾きなどの形状を重要視する距離尺度，異常値などを無視する編集距離に基づく距離尺度は，それぞれ得意なデータセットが異なっていると考えられる．以降の実験では，それぞれのデータセットに対して DDTW，EDR，wDTW のうち，どの距離尺度が正確に類似性を評価できるのかを選択することを考える．

4.2 Locah Shapelet の抽出

Local Shapelet を抽出するためのデータの準備を行う．まず，各データセットに対して最も性能が良い距離尺度を DDTW，EDR，wDTW の中から選択する．そして，選択された距離尺度を，それぞれのデータセットの全ての訓練データにラベルとして付与する．以上のようにして得られたデータから Local Shapelet を抽出する．本実験では，Local Shapelet の部分時系列の長さを 15，$x = 0.9$，$y = 10$ として実験を行った．

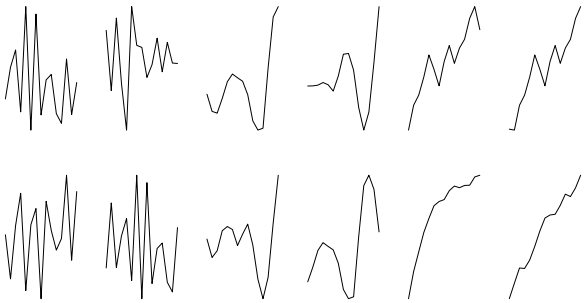

図 2 抽出された Local Shapelet
Fig. 2 Extracted Local Shapelet.

図 2 に本実験で抽出された Local Shapelet の一部を示す．左の 4 つが wDTW が得意なデータセットから抽出された Local Shapelet，中央の 4 つが EDR が得意なデータセットから抽出された Local Shapelet，右の 4 つが DDTW が得意なデータセットから抽出された Local Shapelet である．これらを比較すると，DDTW が得意なデータセットの値の変化はなだらかであることが明らかに分かる．よって，データの値の変化がなだらかなデータセットに対しては，DDTW を用いるのが良いと考えられる．DTW が得意なデータセットの Local Shapelet と，EDR が得意なデータセットの Local Shapelet を比較すると，共に値の変化が激しいことが分かる．しかし，DTW が得意なデータセットの方は細かい変動が多く，EDR が得意なデータセットは大きな変化を含んでいる．このように，各距離尺度が得意なデータセットの特徴は互いに異なり，ある距離尺度が得意な複数のデータセット間では似た特徴を持っていることが分かる．

4.3 データセットに最適な距離尺度の選択

DDTW，EDR，wDTW それぞれに対して 30 個の Local Shapelet を抽出した．この計 90 個の Local Shapelet と，表 1 中の全てのデータセットのテストデータとの BMD を計算し，多数決に基づく方法と 1-NN による分類を用いて，性能が良いと考えられる距離尺度を選択する．表 3 に，各データセットに対する距離尺度の投票結果を示す．表中の太字で表された距離尺度が，各データセットに対して選択

表 3 距離尺度の投票結果
Table 3 Voting result of distance measure.

Data Set	wDTW	EDR	DDTW
50words	0	**341**	114
Adiac	0	1	**390**
Beef	0	**20**	10
CBF	243	**657**	0
Coffee	1	**21**	6
ECG200	**45**	41	14
FaceFour	32	**1658**	0
FaceAll	0	**88**	0
fish	0	**173**	2
GunPoint	0	**150**	0
Lighting2	**54**	7	0
Lighting7	**61**	2	0
OliveOil	0	**30**	0
OSULeaf	0	2	**240**
Swed.Leaf	9	216	**400**
syn.con.	95	**205**	0
Trace	15	0	**48**
TwoPat.	**4000**	0	0
wafer	197	**5967**	0
yoga	18	811	**2171**

表 4 提案手法と EDR，DTW，DDTW を用いた 1-NN によるエラー率
Table 4 Error rates of selected distance measure, EDR, DTW and DDTW using 1-NN.

Data Set	提案手法	EDR	wDTW	DDTW
50words	**0.271**	**0.271**	0.291	0.291
Adiac	**0.380**	0.457	0.446	**0.380**
Beef	**0.400**	**0.400**	0.583	0.585
CBF	0.013	0.013	**0.006**	0.377
Coffee	**0.160**	**0.160**	0.252	0.177
ECG200	**0.153**	0.211	**0.153**	0.167
FaceFour	**0.045**	**0.045**	0.164	0.048
FaceAll	**0.050**	**0.050**	0.079	0.125
fish	**0.090**	0.107	0.261	**0.090**
GunPoint	0.079	0.079	0.055	**0.006**
Lighting2	0.320	**0.088**	0.320	0.273
Lighting7	0.202	**0.093**	0.202	0.431
OliveOil	**0.062**	**0.062**	0.118	0.262
OSULeaf	**0.082**	0.115	0.424	**0.082**
Swed.Leaf	**0.088**	0.145	0.221	**0.088**
syn.con.	0.118	0.118	**0.014**	0.430
Trace	**0.000**	0.150	0.075	**0.000**
TwoPat.	**0.000**	0.001	**0.000**	**0.000**
wafer	**0.002**	**0.002**	0.005	0.010
yoga	**0.078**	0.112	0.151	**0.078**
average	**0.132**	0.134	0.191	0.195

される距離尺度となる．たとえば，50words に対して選択されたのは EDR である．表 4 に，提案手法による分類エラー率と，EDR，DTW，DDTW の分類エラー率を示す．太字は各データセットに対して，一番性能が良い手法を表している．提案手法のエラー率は有意水準 5%で wDTW，DDTW よりも性能が良いことが分かる．また，提案手法と EDR のエラー率には有意差は無い．しかし，一番性能が良かったデータセットの数は提案手法が 15 個，EDR が 9 個，DTW が 3 個である．よって，本手法によりデータセットに対して性能が良い距離尺度を選択できたと言える．

5. まとめと今後の課題

本稿では，Local Shapelet を用いて，データセットに対して性能が良い距離尺度を選択する枠組みを提案した．Local Shapelet は各距離尺度が得意なデータセットの特徴を表すことができる．また，どのような部分時系列を含むデータがどの距離尺度が得意なのかを示した．さらに，実験では既存手法よりも多くのデータセットに対して，正確に類似性を評価できることを示した．一方，H. Ding[1] らは，それぞれの距離尺度に有意な性能差がないと結論づけている．有意な性能差が存在する分類を行うために，本稿では，距離尺度の選択を行ったが，アンサンブル学習などの距離尺度の組み合わせなども考えられる．今後，これらの手法全体に適用できるような枠組みを提案したいと思う．また，データセットの特徴として，ピリオドグラムや自己相関係数など，信号解析，統計に基づいたものを用いて実験て，より良くデータセットの特徴を表現できるものを検討したいと思う．

参考文献

[1] Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. and Keogh, E.: Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures, *Proc. of 34th Very Large Data Bases Endow.*, Vol. 1, No. 2, pp. 1542–1552 (2008).
[2] Keogh, E.: Exact Indexing of Dynamic Time Warping, *Proc. of 28th Very Large Data Bases, 2002*, pp. 406–417 (2002).
[3] Xing, Z., Pei, J., Yu, P. S. and Wang, K.: Extracting Interpretable Features for Early Classification on Time Series, *Proc. of the 11th SIAM International Conference on Data Mining*, pp. 247–258 (2011).
[4] Trotman, A.: Learning to Rank, *Information Retrieval*, Vol. 8, pp. 359–381 (2005).
[5] Keogh, E. and Pazzani, M.: Derivative Dynamic Time Warping, *Proc. of 1st SIAM International Conference on Data Mining*, pp. 1–11 (2001).
[6] Vlachos, M., Gunopoulos, D. and Kollios, G.: Discovering Similar Multidimensional Trajectories, *Proc. of the 18th International Conference on Data Engineering*, pp. 673–684 (2002).
[7] Chen, L., Özsu, M. T. and Oria, V.: Robust and Fast Similarity Search for Moving Object Trajectories, *Proc. of the 2005 ACM SIGMOD International Conference on Management of Data*, pp. 491–502 (2005).
[8] Keogh, E. and Kasetty, S.: On the Need for Time Series Data Mining Benchmarks: a Survey and Empirical Demonstration, *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 102–111 (2002).
[9] Chen, L. and Ng, R.: On The Marriage of Lp-norms and Edit Distance, *Proc. of the 30th International Conference on Very Large Data Bases*, pp. 792–803 (2004).
[10] Assfalg, J., Kriegel, H.-P., Kroger, P., Kunath, P., Pryakhin, A. and Renz, M.: Similarity Search on Time Series Based on Threshold Queries, *Proc. of the 10th International Conference on Advances in Database Technology*, pp. 276–294 (2006).
[11] Chen, Y., Nascimento, M. A., Chin, B., Anthony, O. and Tung, K. H.: Spade: On Shape-based Pattern Detection in Streaming Time Series, *Proc. of the 23rd International Conference on Data Engineering*, pp. 786–795 (2007).
[12] Marzio, M. D.: Kernel Density Classification and Boosting: an L2 Analysis, *Statistics and Computing*, Vol. 15, pp. 113–123 (2005).
[13] saw, J. G., yang, M. C. and mo, T. C.: Chebyshev Inequality with Estimated Mean and Variance, *The American Statistician*, Vol. 38, No. 2, pp. 130–132 (1984).
[14] Keogh, E., Xi, X., Wei, L. and Ratanamahatana, C.: The UCR Time Series Classification/Clustering Homepage [www.cs.ucr.edu/ẽamonn/time_series_data/] (2006).