社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－MPS－73 (11)
2009／3／5

スペクトルと周波数を用いた形状に基づいた時系列データの類似度測定手法

瀧 敬士 † 上原 邦昭 †

† 神戸大学大学院工学研究科

現在，多様な時系列データ間の類似度測定手法が提案されている．しかし，既存の類似度測定手法はデータ点間のユークリッド距離のみに基づき，時系列データの形状を考慮していない．本論文では，時系列データの形状に着目した新しい類似度測定手法 *Distance based on Frequency and Spectrum* (DiFS) を提案する．DiFS は，時系列データを連続したベクトル列から構成される形状と考える．類似度測定に用いる特徴量はベクトルのスペクトル解析により決定する．そして，個々のベクトルのスペクトルを用いて類似度を計測している．さらに，DiFS による類似度測定の計算コスト削減手法を提案する．最終的に，時系列データの評価データを用いた分類実験を行い，提案手法の有効性と効率性を示す．

Distance Measure based on Shape with Frequency and Spectrum in Time Series Data

Keishi TAKI † Kuniaki UEHARA †

† Graduate School of Engineering, Kobe University

In recent years, various metrics have been proposed to calculate similarity between time series data. However most of metrics do not consider shape of time series data. This paper proposes a novel similarity metric, named *Distanced measure based on Frequency and Spectrum* (DiFS), based *not* on individual data points but on shapes of vectors. DiFS calculates similarity between vectors with spectrum. In adddition, we propose efficient method, named eDiFS, utilizing characteristic of DiFS. The effectiveness of DiFS are evaluated through a variety experiments on standard benchmark data sets.

1 まえがき

一般に，人は物体を認識するとき，物体の位置よりも，形状の情報を重要視している．しかしながら，既存の類似度測定手法は位置に依存する手法が多い．たとえば，*Dynamic Time Warping* (DTW) という手法が有名であるが，形状の情報を用いないため，同じ形状の時系列データがあっても．空間的な位置が異なれば，同一であると見なされない．

このため，近年，形状を意識した類似度測定手法が提案されてきている．これらは，時系列データをパターン系列やベクトル系列へと変換して類似性を評価している．しかし，これらの類似度測定関数が形状の評価に適さない場合がある．

また，離散フーリエ変換を用いて，周波数成分のスペクトル間の距離を用いて類似性を定義し，人間の感覚に近い類似性の判定を可能にする手法が提案されている [3]．しかしながら，離散フーリエ変換では，時間領域の情報を失うため，非定常な時系列データには適さないという問題がある．

本論文では，形状に特化した新しい類似度測定手法 DiFS を提案する．DiFS は時系列データをベクトル系列として扱い，連続するベクトル列で時系列データの形状を表現している．さらに，スペクトル解析を行い，人間の感覚に近い，類似性の判定を可能にしている．スペクトル解析では，時間領域を考慮するために，短時間フーリエ変換を用いている．そして，得られたスペクトルを用いた類似度の計測方法を提案している．さらに，動的計画法を用いて、比較する時系列データの長さが異なっても整合を取れるようにしている．

一方，動的計画法は，比較する時系列データ間のデータ点すべての組合せを計算し，最適なパスを求めるため計算コストが高くなる．このため，類似度測定手法の特性を用いた高速化手法が提案されている [4]．たとえば，Morse らは編集距離で用いられている Hunt-Szymanski アルゴリズムを時系列データに応用している．しかしながら，このアルゴリズムはデータ点間の類似度が 0 か 1 の 2

値でしか適用できない．そこで，DiFS はベクトルが同じ向きを向いていないと類似度をとらないという特性を利用して，高速化手法を提案する．最後に，本手法の正当性を評価する実験を行い，類似度測定における効果を示すとともに，効率化についても議論する．

2 Distance based on Frequency and Spectrum

長さ $M+1$ の時系列データ C が次のように表されたとする．

$$C = (t_1, c_1), \ldots, (t_m, c_m), \ldots, (t_{M+1}, c_{M+1}) \quad (1)$$

ここで，c_m は時系列データのデータ点を表しており，t_m はそのデータ点に対応する時間を表している．次に，時系列データ C を，隣り合うデータ点間の差から表されるベクトル系列 $\bar{C}$ へと変換すると

$$\begin{aligned} \bar{C} &= (t_c, \bar{c}_1), \ldots, (t_c, \bar{c}_m), \ldots, (t_c, \bar{c}_M) \\ &= \boldsymbol{c}_1, \ldots, \boldsymbol{c}_m, \ldots, \boldsymbol{c}_M \end{aligned} \quad (2)$$

となり，ここで，時系列データの隣り合うデータ点の時間はサンプリング間隔を表し，一般にサンプリング間隔は一定なので，t_c という定数で与えている．また，$\bar{c}_m = c_m - c_{m-1}$ である．

同様に，ベクトル系列 $\bar{Q}$ は以下のように表されるものとする．

$$\begin{aligned} \bar{Q} &= (t_q, \bar{q}_1), \ldots, (t_q, \bar{q}_n), \ldots, (t_q, \bar{q}_N) \\ &= \boldsymbol{q}_1, \ldots, \boldsymbol{q}_n, \ldots, \boldsymbol{q}_N \end{aligned} \quad (3)$$

本研究では，人の感覚に近い類似性を考慮するために，スペクトルに基づく類似度を定義する．このため，ベクトル列の個々のベクトルに対して，フーリエ変換を行い，スペクトルを抽出する．

ベクトル $\boldsymbol{c}_m = (t_c, \bar{c}_m)$ を例としてスペクトルの抽出について述べる．まず，ベクトルにフーリエ変換を適用するため，ベクトル $\boldsymbol{c}_m$ をすべての実数で定義された関数として 図 1 のような連続波形と考える．

フーリエ変換をベクトル $\boldsymbol{c}_m$ に適用し，そのスペクトル $|F(f)|$ を求めると，

$$|F(f)| = \frac{\bar{c}_m t_c}{f^2} \sqrt{(t_c f - sin(t_c f))^2 + (1 - cos(t_c f))^2} \quad (4)$$

となる．この式をそのまま特徴量として用いると，計算コストが高くなるため，本研究では，周波数

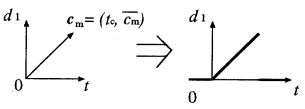

Fig. 1　ベクトルの変換.

が 0 におけるスペクトルを利用する．なお，値は $\lim_{f \to 0} |F(f)| = \bar{c}_m t_c / 2$ と導出することができる．

次に，ベクトル $\boldsymbol{c}_m$ のスペクトルとベクトル $\boldsymbol{q}_n$ のスペクトルを用いてベクトル間の類似度を定義する．類似度は，スペクトル間の比により求める．つまり，時系列データ Q と C のサンプリング間隔 t_q と t_c が等しいとして，ベクトル $\boldsymbol{c}_m$ と $\boldsymbol{q}_n$ 間の類似度は以下のように定義される．

$$\mathrm{sim}(\boldsymbol{q}_n, \boldsymbol{c}_m) = \begin{cases} 1 & \textbf{if } \bar{q}_n = \bar{c}_m = 0 \\ \min\left(\frac{\bar{q}_n}{\bar{c}_m}, \frac{\bar{c}_m}{\bar{q}_n}\right) & \textbf{if } \mathrm{sgn}(\bar{q}_n) = \mathrm{sgn}(\bar{c}_m) \\ 0 & \textbf{otherwize} \end{cases} \quad (5)$$

ここで，$\bar{q}_n = \bar{c}_m = 0$ または $sgn(\bar{q}_n) = sgn(\bar{c}_m)$ ではない場合，つまり，ベクトルの向きが異なれば類似度を 0 と定義している．

以上の定義を用いて，比較する時系列データ中のすべてのベクトルの組合せについて類似度を計算したあと，ベクトル系列 $\bar{Q}$ と $\bar{C}$ 間の類似度は以下の再帰関数を用いて計算される．

$$\begin{aligned} \mathrm{DiFS}(\bar{Q}_n, \bar{C}_m) = \max\{&\mathrm{DiFS}(\bar{Q}_{n-1}, \bar{C}_{m-1}) + \mathrm{sim}(\boldsymbol{q}_n, \boldsymbol{c}_m), \\ &\mathrm{DiFS}(\bar{Q}_{n-1}, \bar{C}_m), \mathrm{DiFS}(\bar{Q}_n, \bar{C}_{m-1})\} \end{aligned} \quad (6)$$

ここで，新しいシンボル $\bar{Q}_n$ と $\bar{C}_m$ はベクトル部分系列 $(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n)$ と $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m)$ を表す．

3 類似度計測の効率化

ベクトル系列 $\bar{Q}$ と $\bar{C}$ から構築された類似度行列の例を図 2 に示す．

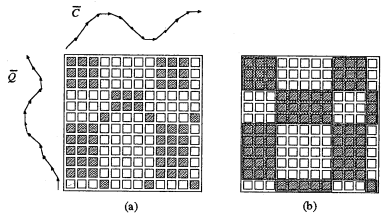

Fig. 2　動的計画法の概要.

図 2 (a) において，白色のセルはベクトル間の類似度が 0，灰色は 0 以外の値であることを示して

いる．通常の類似度計測では，すべてのセルの探索を行う．しかしながら，白色のセル自体は，類似度が0であるため，ベクトル系列間の類似度に影響を及ぼさない．そこで，図2 (b) のような拡大行列内のみで探索を行えば，本来の類似度を変化させず効率的に時系列データ間の類似度を算出することができる．このように，効率的にDiFSの類似度計算を行う手法eDiFSを提案する．

拡大行列を求めるために，類似度が0以外のセルを発見する必要がある．比較するベクトルの向きが同じなら，ベクトル間の類似度は0以外となるため，ベクトル系列の要素をベクトルの向きにより分割すれば，類似度が0となる領域を外して探索することができる．たとえば，図3のベクトル系列 $\bar{C}$ はベクトルの向きが上向きであることを示すインデックス I_1 と，下向きであることを指し示す I_2 に分けることができる．

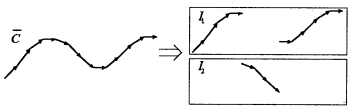

Fig. 3　ベクトル系列のインデックスによる分割．

ここで，比較するベクトル $\boldsymbol{q}_1$ が上向きであれば，インデックス I_1 の要素のみから類似度を算出すればよい．

一方，すべての領域の探索を行わないためには，従来の動的計画法を変更する必要がある．たとえば，白色の領域に隣り合うセルは，白色の領域からのパスを考慮する必要がある．しかし，白色の領域の探索は行わないため，その点に繋がるパスは存在しない．パスの連続性が保持するために．動的計画法を変更する．また，変更例を図4に示す．

図 4 (a) では，cell1 に到達する縦，横方向のパスを表している．式 (6) より，斜めのパスでしか類似度行列内の類似度が加算されることはない．すなわち，横方向だけのパスは，DiFS($\boldsymbol{q}_n, \boldsymbol{c}_m$) =DiFS($\boldsymbol{q}_n, \boldsymbol{c}_{m-3}$) となる．また，縦方向についても，DiFS($\boldsymbol{q}_n, \boldsymbol{c}_m$) =DiFS($\boldsymbol{q}_{n-3}, \boldsymbol{c}_m$) となる．このように，縦，横方向の探索は白色の領域をスキップし，パスの連続性を保持する．

次に，図4 (b) のように，cell1，cell2，cell3に到達する斜め方向のパスについて考える．まず，cell1の斜めのパス p_1 は $p_1 = ((n-1, m-1), (n, m))$ であり，$(n-1, m-1)$ からのパスを作成できるため，連続性が保たれる．次に，cell2の斜めのパス p_2 は，$p_2 = ((n-1, m), (n, m+1))$ である．ここで，$(n-1, m)$ の領域の探索は行わないため，$(n-1, m)$ へ繋がるパスがない．連続性を保持するために，$(n-1, m)$ へ繋がるパスを生成する．

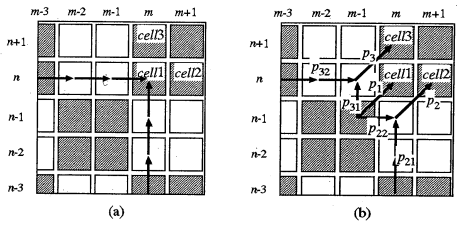

Fig. 4　効率的な探索手法の例．

$(n-1, m)$ に繋がるパスは，図4 (a) のような，縦，横方向のパス p_{21}，p_{22} のみである．つまり，cell2に到達するパスは p_{21} または p_{22} を通り，p_2 を通る必要があるため，$p_{21} + p_2$ と $p_{22} + p_2$ の2本のパスを考慮する必要がある．

同様に，cell3に到達する斜めのパスは $p_{31} + p_3$ と $p_{32} + p_3$ の2本のパスを考慮する必要がある．このように探索を行うことにより，類似度行列をすべて探索することなく，従来の動的計画法と同様の探索を行うことができる．

4　実験

本章では，DiFSの類似度測定手法としての有効性と効率化手法の有効性を確かめる．実験には，時系列データの評価データとしてUCR Time Series Classification/Clustering dataを用いる．

分類の比較として，DTW，また，ベクトル系列のように差分を用いて類似度を算出するDDTW，形状を意識した類似度測定手法であるSpADe，AMSSの分類結果を用いている[1]．

表1は，データセットにおける誤分類率を示し，最も誤差が少なかった値を太字で表している．

表より，DiFSが8データセットで最も正確に分類が行えていることが分かる．これらのデータセッ

[1] DTW，DDTWは我々が実装した実験結果を用い，ワーピング制約は使用していない．なお，他の比較手法の分類結果は論文 2) 1) の値を引用している．

Table 1　分類性能の比較.

Dataset	DTW	DDTW	SpADe	AMSS	DiFS
Gun-Point	0.093	0.007	0.007	**0.000**	**0.000**
Wafer	0.020	0.023	0.012	0.011	**0.000**
OliveOil	0.133	0.200	-	0.200	**0.067**
Yoga	0.164	0.180	0.123	0.158	**0.093**
Coffee	0.179	0.400	-	**0.143**	**0.143**
50 Words	0.310	0.303	0.215	0.242	**0.174**
Beef	0.500	0.467	-	0.433	**0.200**
Adiac	0.396	0.381	0.319	0.345	**0.317**
OSU Leaf	0.409	0.116	0.132	**0.103**	0.140
Swed. Leaf	0.210	0.114	0.125	**0.104**	0.128
Trace	**0.000**	**0.000**	**0.000**	**0.000**	0.050
Fish	0.167	0.103	**0.017**	0.040	0.046
Face (four)	0.170	0.375	**0.034**	0.261	0.205
ECG	0.230	0.170	**0.130**	0.170	0.160
Face (all)	0.192	**0.126**	0.214	0.265	0.227
Two_Pat.	**0.000**	0.002	0.005	0.092	0.147
CBF	**0.003**	0.011	0.020	0.522	0.417
Syn. Con.	**0.007**	0.440	0.080	0.523	0.480
Lightning-2	**0.131**	0.328	0.278	0.180	0.246
Lightning-7	**0.274**	0.425	0.315	0.301	0.493
AVERAGE	0.179	0.217	*NaN*	0.205	0.187

トは比較的滑らかな軌跡であり，形状の評価が容易に行えるためである．一方，CBF，Syn.Con などでは局所的な振動が多く，形状の評価が正しく行えないため分類精度が悪かったと考えられる．

また，誤分類率を AMSS と比較すると，ほとんどのデータセットで本手法の方が有効に分類できていることが分かる．これは，コサイン類似度とスペクトル間類似度の違いによるものである．いい換えると，ベクトル間のコサイン類似度は類似性評価として適切に見えるが，条件によっては，適していないことが分かる．一方，本手法は AMSS と比較して，時系列データの形状にあった類似度であることがいえる．

さらに，SpADe と比較すると，平均的には DiFS のよりも SpADe の性能の方が高い．これは，SpADe が平均振幅を特徴量として用いており，振動の激しいデータに対しても適切に類似性評価ができたためだと考えられる．

図 5 に分類実験に用いたデータのインデックスの構成比率 r と計算効率 E の関係を示す．ここで，E は，eDiFS と DiFS の実行時間の比で表され，eDiFS の実行時間にはインデックス作成にかかる時間も含んでいる．また，図の直線は r と E の関係を近似曲線を用いて表している．

図を見ると，計算効率は計算効率は r に依存していることが分かる．また，近似直線は $E = 1.37r - 0.130$ という式で表される．元の DiFS の計算コストよりも効率的な場合，つまり，$E < 1$ を満たすのは $0.175 < r < 0.825$ の場合となる．r がこの範囲にある場合，eDiFS は効率的に類似度測定を行うことができる．

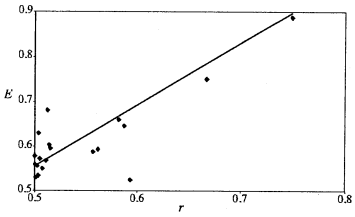

Fig. 5　インデックス構成比率と計算効率の関係.

5　むすび

本論文では，時系列データの形状に基づく類似度測定手法である DiFS を提案した．分類実験では，既存手法よりも形状を妥当に評価できることを示した．eDiFS は DiFS の精度を変えることなく，高速な類似度計測を可能にした．

しかしながら，DiFS にはいくつかの課題がある．まず，振動の激しいデータでは，形状の評価を正しく行えない．そこで，平滑化を組み合わせることにより，的確な形状の評価ができると考えられる．平滑化との組み合わせは別稿で述べる．加えて，サンプリング間隔が異なる場合についても別稿で述べる．また，効率化手法では検索候補数の削減はできないため，検索候補数を削減できる枝狩り手法について考えていく必要がある．

さらに，本研究では単一のベクトルをフーリエ変換したが，複数のベクトルをフーリエ変換する手法，また，周波数が 0 以外のスペクトルを考慮して類似度を定義することもできるため，その点からも類似度測定手法を考案していく必要がある．

参考文献

1) Y. Chen, M. A. Nascimento, B. C. Ooi and A. K. H. SpADe:On Shape-based Pattern Detection in Streaming Time Series, In *ICDE*, pp. 786–795 (2007).

2) 中村哲也，瀧敬士，野宮浩揮，上原邦昭：AMSS:時系列データの効率的な類似度測定手法，電子情報通信学会論文誌，Vol. J91-D, No. 11, pp. 2579–2588 (2008).

3) 小山克正，宝珍輝尚，中西秀哉，小嶋護：視認性を考慮した周波数に基づく時系列データの類似度，技術報告 1，京都工芸繊維大学 (2008).

4) M. D. Morse and J. M. Patel. An Efficient and Accurate Method for Evaluating Time Series Similarity, In *SIGMOD Conference*, pp. 569–580 (2007).