相互情報量に基づく特徴抽出と順序カーネルを利用した時系列データ分類

藤岡 公平† 岡部 正幸†

†県立広島大学 〒734-8558広島県広島市南区宇品東一丁目1番71号

E-mail: †{ r122004ix, okabe} @ed.pu-hiroshima.ac.jp

キーワード Shapelet，時系列データ分類，相互情報量，順序カーネル

1. はじめに

時系列データ分類は，時系列データをその特徴により対応クラスに分類する時系列データ解析技術の一つである．解析対象となるデータは，センサーから得られる心拍・温度・湿度や株価・為替レートといった経済指標など多岐にわたる．また，これらのデータの分類予測に基づき状態推定や行動認識が行われるなど様々に応用されている．時系列データ分類に関する手法はこれまでに様々なものが提案されているが，Shapeletを用いた分類手法は高性能な手法の一つとして知られている．Shapeletとは，時系列データ中の特徴的な部分時系列データの事を表す．ここでいう特徴的とは，ある時系列データを別の時系列データと区別する際に有効であるということを意味する．Shapeletは時系列データの判別モデルを生成する際の特徴集合として用いられるため，Shapelet集合の品質はモデルの予測性能に大きな影響を与える．

Shapeletの抽出手法の一つに，Grabockaらが提案した探索的アプローチによるものがある[1]．この手法は，まず時系列データからShapeletの候補となる部分時系列をランダムに切り出し，切り出した順に1つずつ部分時系列の分類精度の計算を行い，分類精度を向上させるものをShapeletの1つとして抽出する．そして，この処理を候補数分繰り返すことで最終的なShapelet集合の抽出を行うものとなる．しかし，分類精度の計算順序はランダムであるため，繰り返しの序盤に性能の悪いShapeletが採用され，最終的な分類精度に悪影響を与える可能性がある．例として，1回目の計算で分類精度が0.3の場合，精度は初期値の0よりも向上するためShapeletとして抽出されるが，性能が悪いため最終的なShapelet集合の中では足を引っ張るShapeletとなる．そのため，本研究ではShapelet抽出時の精度計算順に着目し，その指標として相互情報量を導入することにより，更なる精度向上を目指す．

また，Shapeletによる時系列データ分類は，Shapeletと各時系列データとの最小距離を計算し，最小距離同士の内積となる最小距離カーネルを特徴としてSVMなどの分類器へ導入することにより行われる．その際，いずれも個々のShapeletは単独の特徴として取り扱われ，Shapelet間の関係性は考慮されない．しかし，時系列データ分類において重要となるデータ間の類似性を計算する際に，特徴となるShapelet間の相対的な順序関係を用いた方が有用である場合がある．例えば，行動認識データにおいて，マウスを移動してクリックする動作とクリックしてから移動させる動作は，部分的な動作(Shapelet)は似ているが，順序を考慮しないと同じ動作として認識されてしまう．そのため，本研究ではShapelet間の相対的な順序関係に着目し，出現順序の類似度を定量化したカーネルを導入することにより，更なる精度向上を目指す．

2. 相互情報量に基づく特徴抽出

分類時の特徴となるShapeletの抽出には，まず従来手法と同様にShapeletの候補となる部分時系列uを時系列データ集合Xからランダムにt個切り出し，部分系列集合Uを生成する．

$$U = \{u_1, \dots, u_t\}$$

次に，部分時系列を１つずつ取り出し，Shapeletと仮定して，分類精度の計算を繰り返し行う．精度の計算には，各部分時系列と各データとの最小距離を特徴として最近傍法を用いて行う．精度が初期値を0として，それ以前の繰り返しでの最高精度以上の場合，取り出した部分時系列をShapelet集合に追加する．本研究ではこの繰り返しの中で取り出す部分時系列の順序を，各部分時系列の持つ相互情報量$I(C;s)$の降順とする．相互情報量とは，2つの確率変数の相互依存の尺度を表す量となり，この値が大きいほどクラスの特徴を表す特徴と見なすことができる．そのため，相互情報量の降順での計算は，分類への寄与度の大きな部分時系列からの計算が可能になり，性能の悪いShapeletを抽出することを抑制することに繋がる．

$$\begin{aligned} I(C;s) &= H(C) - H(C|s) \\ &= \sum_{c \in C} \sum_{f_t \in 0,1} P(c, f_s) \log\left(\frac{P(c, f_s)}{P(c)P(f_s)}\right) \end{aligned}$$

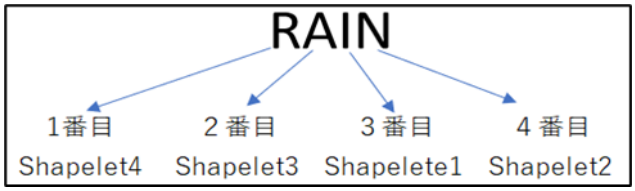


図1 文字列と出現したShapeletの対応の例

ここで，Hはエントロピー，Nは全データ数，f_sは部分時系列sがあるデータに出現したかどうかを表しており，$P(c, f_s = 1)$と$P(c, f_s = 0)$は部分時系列sがクラスcのデータに出現する確率と出現しない確率，$P(c)$はあるデータがクラスに所属する確率，$P(f_s = 1)$と$P(f_s = 0)$は部分時系列sが出現する確率を表している．

また，各データにおいて各部分時系列が出現したかどうかの判定には，まず，ある部分時系列u_tとあるデータX_nとの最小距離$d_{n,t}$を計算する．

$$d_{n,t} = \min_c \left(X_{n,c} - u_t\right)^2$$

ここで，$X_{n,c}$はn番目の時系列データX_nをu_tと同じ長さで切り出したc番目の部分系列である．次に$d_{n,t}$全ての値に対して昇順にソートし，先頭からp(%)の最小距離を閾値dpとする．$d_{n,t}$の値が閾値以下であれば部分時系列u_tがデータnにおいて出現したと判定する．また，pはハイパーパラメータとなる．

$$f_t = \begin{cases} 1 & if \quad d_{n,t} \le dp \\ 0 & else \ \ d_{n,t} > dp \end{cases}$$

このように閾値を用いることで，出現判定の計算が可能となる．

3. 順序カーネルを利用した時系列データ分類

Shapletの抽出には，前節で説明した提案手法を用いて抽出を行う．これにより，Shapelet集合Sと，各ShapeletとデータX_nとの最小距離を要素として持つ最小距離ベクトルf_n^{md}を得る．最小距離計算は，相互情報量の計算の際に完了しており，その結果を用いてベクトルを生成する．また，Shapelet集合は，複数のShapelet長毎にShapelet集合を作成し，それらすべてを含んだ集合となる．

$$S = (s_1, s_2, \dots, s_i, \dots, s_m)$$
$$f_n^{md} = \left(d_{n,1}, d_{n,2}, .., d_{n,i}, \dots, d_{n,m}\right)$$

本研究では更に，Shapelet間の順序関係に基づく特徴を導入する．文字列データ間の類似度を計算する手法として，n-gramカーネルがある[2]．本研究では，図1のように文字列データにおける各文字をShapelet

表1 実験で使用したデータセットの詳細

データセット名	訓練データ数	テストデータ数	長さ	クラス数
ECG200	100	100	96	2
GunPoint	50	150	150	2
Lightning2	60	61	637	2
MedicalImages	381	760	99	10
SonyAIBORobotSurface1	20	601	70	2
Yoga	300	3000	426	2

に見立て，Shapelet版のn-gramカーネルを生成し，これを順序カーネルと呼ぶ．この時，文字と見立てるShapeletは時系列データ毎に異なり，時系列データ上に出現するShapeletのみを文字と見立て，出現するShapeletのみでの順序を考える．出現判定には，Shapelet抽出時の相互情報量の計算内で，各データにおける各部分時系列が出現したかどうかの判定を行う際の閾値を使用し，Shapeletと時系列データとの最小距離が閾値以下のものを出現したとする．

今，時系列データXとYがあるとする．順序カーネルを計算するにあたり，まず，XとYそれぞれにおける各Shapeletの出現位置を，最小距離を用いて求め，出現順にShapeletを並べた系列X^s, Y^sに変換する．

この系列X^s, Y^sを利用して，長さnのn-gram出現頻度ベクトルf^{ng}を求める．ここで，n-gram出現頻度ベクトルとは，あるデータ上に出現するShapelet順でShapeletを並べた系列内に，n個のShapeletの並びの系列の集合L={l_1, …, $l_{{}_{NP}{}_n}$}の各要素が出現するか否かを保存したベクトルで，要素の値は出現する場合は1，出現しない場合は0となる．また，ベクトルの次元数は，Shapeletの要素数Nから生成可能な長さnの順列数${}_NP_n$である．

$$f^{ng}(X^s) = (w_1, \dots, w_{{}_{NP}{}_n})$$

ここで，

$$w_i = \begin{cases} 1 & (l_i \ appears \ in \ X^s) \\ 0 & (otherwise) \end{cases}$$

である．

X^s, Y^sから生成されたn-gram出現頻度ベクトルをそれぞれ$f^{ng}(X^s)$, $f^{ng}(Y^s)$とすると，順序カーネルK^{ng}の各値は$f^{ng}(X^s)$, $f^{ng}(Y^s)$の内積によって計算できる．

$$K^{ng}(X^s, Y^s) = \langle f^{ng}(X^s), f^{ng}(Y^s) \rangle$$

これにより，各時系列データにおけるShapeletの出現順序に関する類似性を求めることとした．本研究では，最小距離ベクトルを合わせて利用するため，XとYの最小距離ベクトル$f^{md}(X)$, $f^{md}(Y)$の内積によって求まる最小距離カーネルK^{md}を求める．

$$K^{md} = \langle f^{md}(X), f^{md}(Y) \rangle$$

最終的な分類は，上記の順序カーネルK^{ng}と最小距離カーネルK^{md}を加算したマルチカーネルK^{mul}をSVMへ導入することにより行われる．

$$K^{mul} = K^{ng} + K^{md}$$

4. 実験

提案手法の性能を評価するため，時系列データを分類する4つの従来手法と比較する実験を行った,評価尺度には精度を用いた．使用したデータセットは22個でそのうちの6個を表1に示す．提案手法はShapeletの抽出とShapeletを用いた分類の双方を用い，Shapelet長はデータ長の10%,20%,30%とし,切り出す部分時系列数は1000個,分類器はSVMとした．提案手法は，相互情報量に基づくShapelet抽出の効果と順序カーネルの効果をそれぞれ調べるために，2通りとした．1つは順序カーネルと最小距離カーネルの両方用いた分類の提案手法①,もう1つは最小距離カーネルのみを用いた分類の提案手法②とし，どちらにおいても相互情報量に基づくShapelet抽出手法を使用した．比較する従来手法はSD，ST，FS，BOSSの4つとなる．SDはGrabockaらが提案した，探索的アプローチによるShapeletを用いた分類手法となる.STはHillsらが提案したShapeletベースの分類手法で，Shapelet候補となる部分時系列をすべて切り出し，評価の高い部分時系列をShapeletとして用いて分類する手法となる[3].FSはRakthanmanonらが提案したShapeletベースの分類手法で,部分時系列をSAXワードと呼ばれるアルファベットの並びに変換し,SAXワードの持つ識別力が高い部分時系列をShapeletとして用いて分類する手法となる[4]．BOSSはSchäferらが提案した分類手法で，時系列データをウィンドウ毎に離散フーリエ変換し，Multiple Coefficient Binningを使用して，単語を作成し，単語の出現頻度のヒストグラムを特徴として分類を行う手法となる[5].

結果を表2に示す．提案①の列は，順序カーネルと最小距離カーネルの両方を用いた分類で，提案②の列は，最小距離カーネルのみを用いた分類となる．提案①の22個のデータセットにおける精度の平均順位は6手法中2.2位，提案②は2.9位，SDは5.1位，STは2.5位，FSは5.5位，BOSSは2.6位となり，今回提案した手法すべてを用いた①の平均順位は，他の手法と比較して最高順位となった．提案するShapelet抽出手法のみを使用した提案②と提案元となったSDとの順位を比較すると，提案②の方が順位が高い．これにより，Shapelet抽出時の評価順序がランダムであることにより，性能の悪いShapeletが抽出されるという問題点が改善され

表2　従来手法との比較

	提案①	提案②	SD	ST	FS	BOSS
ECG200	**0.8850**	0.8530	0.8180	0.8300	0.7660	0.8700
GunPoint	0.9853	0.9524	0.9310	**1.0000**	0.9330	**1.0000**
Lightning2	0.8164	0.8164	0.7950	0.7377	0.7070	**0.8361**
MedicalImages	**0.7496**	0.7074	0.6760	0.6697	0.5960	0.7184
SonyAIBORobotSurface1	**0.9647**	0.9622	0.8500	0.8436	0.6860	0.6323
Yoga	0.8331	0.8223	0.6250	0.8177	0.7050	**0.9183**

表3 提案手法①と②の比較

データ	順序ｶｰﾈﾙ	精度	±	標準偏差
ECG200	なし	0.8530	±	0.0142
	n=2	0.8790	±	0.0130
	n=3	0.8830	±	0.0135
	n=4	0.8840	±	0.0143
p=0.9	n=5	**0.8850**	±	0.0143
GunPoint	なし	0.9524	±	0.0303
	n=2	**0.9853**	±	0.0040
	n=3	0.9840	±	0.0053
	n=4	0.9840	±	0.0053
p=0.2	n=5	0.9840	±	0.0053
Lightning 2	なし	0.7849	±	0.0221
	n=2	**0.7863**	±	0.0205
	n=3	0.7863	±	0.0205
	n=4	0.7863	±	0.0205
p=0.1	n=5	0.7863	±	0.0205
Medical Images	なし	0.7074	±	0.0112
	n=2	**0.7496**	±	0.0077
	n=3	0.7486	±	0.0069
	n=4	0.7493	±	0.0055
p=0.9	n=5	0.7484	±	0.0067
SonyAIBO Robot Surface1	なし	0.9622	±	0.0130
	n=2	0.9637	±	0.0122
	n=3	**0.9647**	±	0.0123
	n=4	0.9646	±	0.0128
p=0.2	n=5	0.9647	±	0.0128
Yoga	なし	0.8223	±	0.0182
	n=2	0.8267	±	0.0221
	n=3	0.8317	±	0.0216
	n=4	0.8317	±	0.0216
p=0.8	n=5	**0.8331**	±	0.0190

提案手法の有効性が示された．また，提案した分類手法を使用した提案①と使用しなかった提案②を比較すると，提案①の方が順位が高い．そのため，Shapeletの順序関係に基づく特徴の有効性が示された．

また，順序カーネルを使用しない場合の精度と，順序カーネルで保存するShapeletの並びに含まれるShapeletの個数nの値を2～5に変化させたときの精度を表3に示す.「順序カーネル」の列の「なし」は，提案手法②，n=2～5は提案①における順序カーネルのnの値を2～5に変更した結果となる．

実験結果より，順序情報を特徴として使用した場合と使用しなかった場合を比較した時，22データセット中19データセットが精度向上かつ標準偏差の縮小という結果となった．これにより，提案手法の有効性が示され，Shapeletの出現順序という情報は分類に有用であると言える．

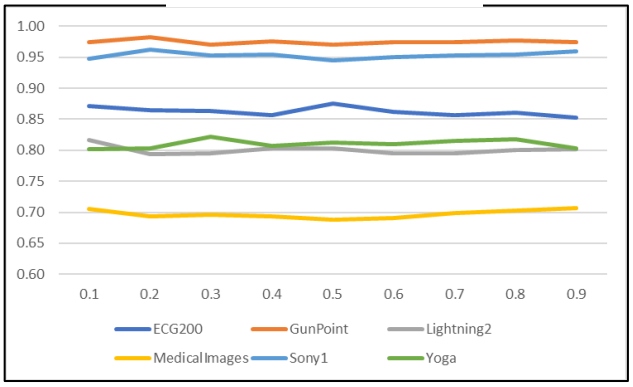
5. 考察

ここでは，順序カーネルで保存されるn個のShapeletを並べた系列の変化による影響を考察する．順序カーネルで保存されるnのShapeletを並べた系列のnの値をn+1に増加させると，n+1の場合にデータ間で一致するShapeletの並びの系列の数は，nの場合にデータ間で一致した系列数以下となる．例として，5つのShapeletがデータ1において(2, 4, 1, 3, 5)，データ2において(2, 4, 1, 5, 3)の順序で出現した場合を考える．n=3の場合に含まれるShapeletの並びの系列はデータ1においては(2, 4, 1)と(4, 1, 3)と(1, 3, 5)の3つ，データ2においては，(2, 4, 1)と(4, 1, 5)と(1, 5, 3)の3つとなり，データ間で一致する系列は(2, 4, 1)の1つとなる．しかし，nを1増やしたn=4の場合に含まれるShapeletの並びの系列はデータ1においては(2, 4, 1, 3)と(4, 1, 3, 5)の2つ，データ2においては(2, 4, 1, 5)と(4, 1, 3, 5)の2つとなり，データ間で一致する系列数は0個で，n=3の場合よりも一致数が減少している．この原因は，保存する系列長を長くしたことにより，n-gram出現頻度ベクトルで保存されるShapeletの出現した時の系列数が減少することと，系列長が短いときは一致していたが，長くすると一致しなくなる系列が出現することである．そのため，nを増加した時に，一致系列数は減少すると考えられる．しかしn=2からn=3にかけて精度が向上している表5のSonyAIBORobotSurface1を見ると，nを増加した場合においても精度が向上しているため，系列１つあたりの順序関係の情報がもたらす効果は増加している．このことから，nの値が増加すると，2つのデータ間における系列の一致数は減るが，系列長が長い分だけモデルの分類性能に対する寄与度が大きいと考えられる．ここで，寄与度とは影響度合いを意味する．

次に精度が最大となったn=3以降の精度を見てみると，n=4では一度精度が低下し，n=5おいて再び向上している．これは，出現したShapeletを並べた系列が一致する数がn=3の時と比較してn=4の時は減りすぎてしまい，nが増加し系列長が長くなったことによる分類への寄与度の増加よりも，一致する系列数の減少による悪影響の方が大きいためと考えられる．一方n=4から5への増加は一致する系列数の減少よりも，nの増加による分類への寄与度の増加の方が大きいため，全体として分類へのプラス要因の方が大きくなり精度の向上につながったと考えられる．このことから，Shapeletの順序関係による時系列データの分類には，nの増加による一致系列数と系列長による分類への寄与度がトレードオフのように関係していると考えられる．そのため，nの値を適切に設定することが重要と言える．

次にShapelet抽出時の相互情報量の計算内で，Shapeletの出現判定の閾値に使用するpの変化による

表5 pの値の変化による影響



精度への影響を調べるために，表5に棒グラフで提案した抽出方法のみを使用した時の精度の変化を示した．横軸がpの値，縦軸が精度となっている．データセット全体で見てみると，いずれのデータセットにおいても，pが変化したことによる精度の変化は起きているが，その変化の仕方はそれぞれで異なっている．そのため，最適なpはデータセットごとに異なっていると言える．

6. まとめ

本研究では時系列データ分類におけるShapeletの抽出手法と，Shapeletを使用した分類手法を提案した．Shapeletの抽出手法では性能の悪いShapeletが抽出されることを防ぐため，部分時系列がShapeletとして有用かどうか評価を行う際の精度の計算順を，相互情報量の降順として計算することを提案した．分類手法ではShapeletの順序性を考慮するために，文字列データにおける各文字をShapeletに見立て，順序カーネルを用いた時系列データ分類の分類手法を提案した．データセットを用いて実験を行った結果，精度が向上し，提案手法の有効性を確認することができた．

参 考 文 献

[1] J. Grabocka, M. Wistuba and L. Schmidt-Thieme, “Fast classification of univariate and multivariate time series through Shapelet discovery,” Knowl Inf Syst 49, pp.429-454, 2016.

[2] H. Lodhi , C. Saunders, J. Shawe-Taylor, N. Cristianini,and C. Watkins,“Text classification using string kernels”, Journal of Machine Learning Research, Vol. 2, pp. 419-444, 2002.

[3] J. Hills, J. Lines, E. Baranauskas, J. Mapp, A. Bagnall, “Classification of time series by shapelet transformation”, Data Mining and Knowledge Discovery volume 28, pp. 851–881 , 2014.

[4] T. Rakthanmanon and E.Keogh, “Fast Shapelets: A Scalable Algorithm for Discovering Time Series Shapelets”, Procee-dings of the 13th SIAM international confere-nce on data mining, 2013.

[5] P. Schäfer, “The BOSS is concerned with time series classification in the presence of noise”, Data Mining and Knowledge Discovery volume 29, pp. 1505–1530, 2015.