Vol.2014-MPS-97 No.3
2014/3/3

時系列分類のための部分類型に基づく特徴量

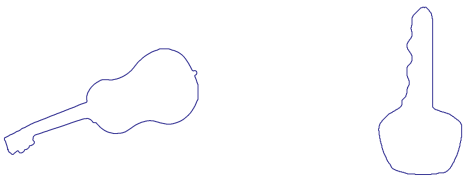
嶌 真由理[1,a)]　安藤 普[2,b)]　関 庸一[2,c)]

概要：時系列分類問題は機械学習の重要な応用の一つである．しかし，時系列データは順序構造を持つベクトルであるため，直接分類器の入力とすることには適さない．このため分類学習に適した特徴量を生成することが重要な課題となっている．従来研究における有望なアプローチの一つに，時系列内に頻繁に現れる部分形状を特徴量として用いる方法がある．しかし，部分形状選択の効率化のため基準として情報利得を用いることから訓練集合の一部にしか現れない部分形状を抽出できないという問題があった．これに対し，本研究では時系列を部分形状の集合とみなす多事例学習 (Multiple Instance Learning) の考え方を利用した特徴量生成の方法を提案する．提案手法ではまず，部分系列の階層的クラスタリングに基づきクラス固有な部分類型を生成する．さらに，集合間の距離尺度を用いて各類型と時系列の距離を評価し，時系列を表す特徴ベクトルとして定義する．訓練集合の一部にしか現れない部分形状も利用して時系列を特徴付けるため，従来手法よりもロバストな分類を実現する．数値実験では，3 つのベンチマークを用いて提案手法の性能を評価し，既存手法との比較を示す．

1. はじめに

実世界事象の観測結果は多くの場合系列の形をとる．このため系列データの分類は実用的に重要な応用問題である．系列データでは連続して観測される値が相関する構造を持つ．例えば，時系列データでは時間軸上の隣接する観測値間に相関が存在する．また，図 1 のような 2 値画像では輪郭の隣接する点間に相関が存在する．系列データはこのような構造から一般的な分類器の入力として直接扱うことは難しいことが知られている．

一方，人は系列データに関しては高い判別能力を持つ．部分的な形状の特徴を認識し，さらに既存の知識と照合することで判別を行うと考えられる．図 1 は輪郭から画像のクラスを判別する問題である．(a) がギター，(b) が鍵という輪郭が類似した難しい問題であるが，この例ではギターのネックと胴体の部分，鍵の溝と握りの部分等を既存知識に基づいて認識し，さらに画像全体のクラスを判別することができる．

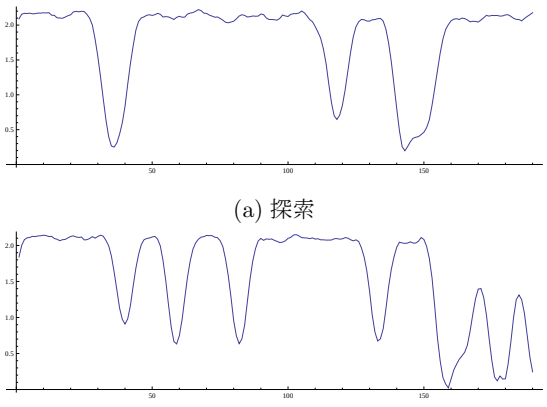


(a) ギター　(b) 鍵

図 1: 輪郭データの例

また，図 2(a),(b) は自律エージェントの速度プロファイルを示したものである．縦軸は速度の絶対値，横軸は単位時間である．それぞれは異なるタスクを実行している際のプロファイルであり，追跡は細かく上下に変化している所に特徴的な部分形状が見られる．しかし，一定の速度の部分は二つのタスク間で共通なパターンである．このような場合にも人はタスクに特徴的な部分的形状を認識し，クラスの違いを判別することができる．

人は上に述べたように部分形状を高速に認識することが



(a) 探索　(b) 追跡

図 2: 速度プロファイルの例

1 群馬大学大学院工学研究科
2 群馬大学理工学研究院
a) t12801713@gunma-u.ac.jp
b) ando@cs.gunma-u.ac.jp
c) seki@cs.gunma-u.ac.jp

可能であるが，機械学習においてはそのような学習は効果的では無い．さらに，系列を部分形状に分割した際に，クラス間で共通する形状が多数生成する．訓練データのラベルとして与えられる教師情報から分類に役立つ特徴的な形状を効率的に抽出し，分類器の学習に効果的に利用することが課題となっている．

時系列データの判別学習においては部分形状を用いるという点で人の認識に類似した方法が効果的であることが示されている [13][7][3]．最も代表的な手法は，Shapelet[9] と呼ばれる特徴的な部分形状が時系列内での検出・非検出を属性とする分類木を構築し，判別を行っている．

Shapelet は訓練時系列データを類似度に基づいて分割した場合に最大の情報利得が得られる部分形状と定義される．この定義から，Shapelet 手法における目標は，出来るだけ多くの訓練データに見られる少数の部分形状を抽出することといえる．すなわち，少数の訓練事例でしか見られない部分形状はクラスに特徴的であっても情報利得は高くならない．図 1，2 のようにクラス間で共通する形状が多い場合やクラスを表す特徴的な形状が複数ある場合には異なる方法論が必要と考えられる．

本研究では上記の議論から時系列が異なるクラスを持つ部分系列事例の集合とみなし，多事例学習 [11] の考え方に基づく時系列分類手法を提案する．提案手法ではまず，[13] による教師付きクラスタリング手法を部分形状の集合に適用することで，ラベルが一様なクラス内クラスタを得る．これをクラスに特徴的な部分形状を表すものと考え，時系列とクラスタ間の距離を時系列の特徴量として定義する．時系列とクラスタはともに部分形状の集合である事から距離関数としてハウスドルフ距離を用いる．ハウスドルフ距離は多事例学習において事例ベース学習に用いられる集合間の距離を表す尺度である．

提案手法では各クラスタへのハウスドルフ距離のベクトルを特徴量としてをサポートベクトルマシンを学習する．従来手法と比較した際の利点は，クラスタリングを利用することで複数の特徴的形状を抽出でき，さらに，クラスタと類似度の高い部分形状がクラス間で共通してみられる場合にも特徴量に違いが生じない点である．

本研究では提案手法の性能を人工時系列データと実世界時系列データを用いた数値実験において評価した．また，実世界データにおいて抽出された特徴部分形状を視覚的に分析した．

2. 関連研究

2.1 Shapelet 手法

一般的な分類手法は，独立な変数値のベクトルを入力とするため時系列のような構造を持つデータを扱うことが難しい．対処するための有効な手段の一つとして時系列を特徴づける有益な部分系列を認識し，分類学習に利用する方法が提案されている．代表的なものが [9] において提案された時系列 Shapelet を用いる手法である．Shapelet は時系列の部分系列と閾値の対で構成され，訓練例の時系列集合を自身との類似度・距離によって分割した際に最大の情報利得が得られるものと定義される．

Shapelet は時系列の視覚的な分析に利用される他，分類モデルの構築にも利用される．[10] では Shapelet との一致を閾値により分類し，その結果を変数として決定木を構築している．[14] では，Shapelet との一致に基づいて時系列を分類する Local Shapelet と呼ばれる手法が提案されている．[9] では情報利得や F 比などの評価尺度が高い複数の Shapelets を抽出し，各 Shapelet との距離を特徴ベクトルとして時系列を表現する Shapelet transformation が提案されている．

Shapelet は情報利得やそれに類する単一の基準に基づいて選択されるため，非常に高速なアルゴリズムにより抽出することが可能である．同時に，いくつかの問題点も指摘されている．個々の Shapelet は別々に評価されるため，Shapelet 間の依存関係をふまえて選択することは難しい．また，クラスの小さい割合の事例のみに含まれる部分系列は情報利得が小さくなるため，それが固有のパターンである場合でも見落とされやすい．これは，部分系列の大部分がクラス間で共通するパターンを持つ場合には重大な問題となる．上に挙げた問題は特徴選択に関する議論と同様である．これは Shapelet が判別のための情報を提供する役割を持つことからくる．

2.2 多事例学習

多事例学習では事例が属する集合が与えられている特殊な設定を扱う．通常の教師あり学習では各事例にラベルが与えられるが，多事例学習では事例の集合にのみラベルが与えられる．多事例学習は画像分類やウェブマイニングなど，多くの分野で適用されている [11]．

本研究では，時系列を部分系列の集合とみなし，さらに部分系列は時系列のクラスに固有のパターンとクラス間共通のパターンが混在するケースを扱う．これは多事例学習の前提に近いたため，その考え方を導入することで従来の Shapelet で扱うのが困難な問題の解決を図る．具体的には多事例学習で用いられるハウスドルフ距離関数を時系列・部分系列間の距離として用い，時系列の特徴付けに利用する．

2.3 類型パターンに基づく最近傍法

時系列分類手法の中では最近傍法がロバストであることが知られている [8]．また，[13] では時系列の訓練集合がクリーンでない場合に対応するため，時系列のクラスタ類型を最近傍法の訓練事例として用いる手法が提案されている．この手法では教師付き階層的クラスタリングを行い，

単一クラスの事例から構成されるクラスクラスタを抽出する．その事例を最近傍法の訓練例とすることで，複数クラスが重複する領域の事例が除外されるため，最近傍法の予測を汎化する効果が得られる．

3. 準備

3.1 一様クラスタ

[13] では次のような教師付き階層クラスタリングから得られる部分木を一様クラスタと呼ぶ．

(1) 部分系列の集合を入力とし，単連結クラスタリングにより階層木構造を作成する．
(2) 部分木のうち，葉の数が指定された数 m 以上でかつすべての葉のラベルが同じものを抽出する．これを一様クラスタと呼ぶ．

一様クラスタの要素を訓練例として用いることで最近傍法を汎化する効果が得られる．

4. 提案手法

提案手法では訓練データからクラスに特徴的な部分形状をクラスタリングにより抽出し，さらに各データについて部分形状との類似度に基づく特徴量を生成する．類似度を求める際に，多事例学習の考え方に基づいて一つの時系列を部分形状を含む集合として扱う．以下にその手順を示す．

時刻 t における観測値を x_t とする． i 番目の観測時系列を X_i とし，時刻 $1,\cdots,T$ における観測時系列を $X=(x_1,\cdots,x_T)$ と表記する．$\mathcal{X}=\{X_i\}_{i=1}^n$ を訓練データとし，ラベルを $\mathcal{Y}=\{y_i\}_{i=1}^n$ と表記する．X_i を長さ l の部分系列に分割する．各部分系列は $\mathbf{s}_t=(x_i,\cdots,x_{i+l-1})$ と表され，X_i の部分系列集合 S_i は $S_i=\cup_{i=1}^n \mathbf{s}_i$ と表される．全訓練データの部分系列の和集合を $\mathcal{S}=\{S_1,\cdots,S_n\}$ と記す．

$\mathcal{S}$ を入力とし，一様階層クラスタリング [1] により部分系列クラスタの集合 $\mathcal{C}=\{C_j\}_{j=1}^q$ を生成する．

ハウスドルフ距離は時系列とクラスタの間の距離尺度として導入する．ユークリッド距離を D としたとき，時系列 S_i とクラスタ C_j のハウスドルフ距離は次式で与えられる．

$$d(S_i,C_j)=\max_{C\in C_j},\min_{S\in S_i}\{D(C,S)\} \quad (1)$$

X_i の特徴量ベクトル $\mathbf{v}(X_i)$ を次のように定義する．

$$\mathbf{v}(X_i)=(d(S_i,C_1),\cdots,d(S_i,C_q)) \quad (2)$$

本研究では $\{\mathbf{v}(X_i)\}$ を特徴量として長さ q の重みベクトル $\mathbf{w}$ をパラメータとする線形分類器 $f:\mathbb{R}^q\to\{1,-1\}$ を学習する．

$$f(\mathbf{v})=\mathrm{sgn}(\mathbf{wv}) \quad (3)$$

$\{\mathbf{v}X_i\}$ を X_i のハウスドルフテンプレート変換 (Hausdorff Template Transform:HaTT) と呼ぶ．テストデータについても (1)(2) 式から HaTT を生成し，学習した f を用いて分類を行う．

5. 実験

5.1 データ概要

本研究では人工データと実世界データそれぞれ二つを用いた実験により，提案手法の検証を行った．人工データは時系列分類の標準的ベンチマークである Synthetic Control Chart[4] と Cylinder-Bell-Funnel 関数 [5] を用いる．

実世界データの概要を以下に述べる．

5.1.1 輪郭データ

MPEG7 CE Shape-1 は，2 値画像の集合で MPEG7 インターフェイスの形状記述子のためのテストデータである [6]．[3] では画像から抽出した輪郭を時系列に変換し，時系列分類手法を適用している．本実験では分類が難しいとされる [6] ギターと鍵の画像から 2 クラス分類問題を用意する．輪郭から時系列を生成する手順は [3] にならった．図 1(a)(b) はそれぞれ guitar，keys の画像の例である．

5.1.2 速度プロファイルデータ

マルチエージェント実験 [12] における 2 台の自律エージェントの速度プロファイルからその行動を分類する問題である．エージェントの行動は探索と追跡の 2 つがある．図 2(a)(b) に示した時系列はそれぞれ追跡，探索のプロファイルの例である．

5.2 設定

前節に示した各データから分類問題を設定する．データの概要を表 1 に示す．

提案手法のパラメータ m はそれぞれの問題で $m=30,40,60,50$ とする．部分系列の長さ w はそれぞれの問題で $w=40,100,100,100$ とする．線形 SVM の学習には LibSVM[*1][1] を用いる[*2]．

比較手法として最近傍法，Local Shapelet[14]，HMM 分類器を用いる．最近傍法は速度プロファイルおよび輪郭データに自然に適応する事が難しいため，適用を行わない．速度プロファイルではパターンの長さが一定ではないためで，輪郭データでは画像間で明確に対応する点を決定する事が難しいためである．

隠れマルコフモデル (HMM) は時系列の生成モデルとし

表 1: 各データ概要

	事例数	クラス数	分割比	時系列長
control	6	300	1 : 1	60
CBF	3	300	1 : 1	129
輪郭	2	20	4 : 1	500
速度	2	4	1 : 1	1996

[*1] www.csie.ntu.edu.tw/˜cjlin/libsvm/
[*2] LibSVM のパラメータは-t 0 -s 0 -c 10 とする．

て広く用いられる．潜在変数が離散変数であり，かつ変数の状態遷移がマルコフ過程であると仮定したモデルである [2]．一つの時刻について見ると，成分密度分布が過去の観測で選択された成分に依存して選択されるように混合分布モデルを拡張したものである．HMM を用いた分類は音声認識や自然言語分析などに広く利用されている．本研究では HMMWeka[*3] の実装による HMM 分類器を用いる．

最近傍法の近傍数は $k = 3$ とする．Local Shapelet のパラメータは $p = 0.8$ とする．輪郭データはクラス数と時系列長が少ないため交差検定を行った．交差検定におけるパラメータ k は $k = 5$ である．

5.3 結果

提案手法および比較手法の率を表 2 に示す．太字は最も高い値を示す．

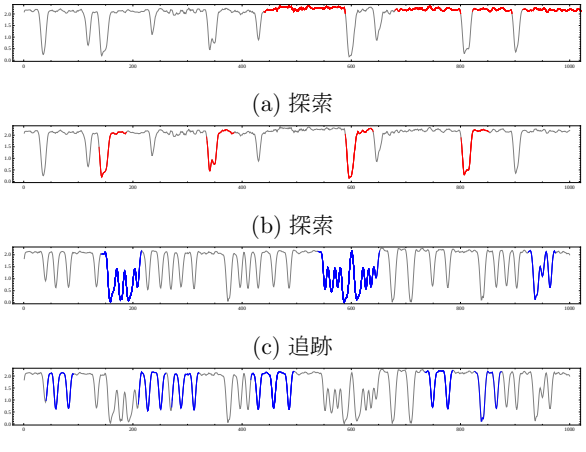
提案手法は CBF を除いて最も高い正解率を示した．CBF においては HMM が最も高い正解率を示し，提案手法もそれに近い正解率を達成した．速度プロファイル，輪郭データの訓練例から一様クラスタとして抽出した部分系列の典型例をそれぞれ図 3，4 に示す．

図 3 の形状がそれぞれのクラスを特徴付ける形状といえるのは，探索では (a) のような一定の速度となっている形状や，(b) のような急激に速度が変化している形状が挙げられる．追跡では (c),(d) のように速度が小刻みに変化する形状が挙げられている．

図 4 の形状がそれぞれのクラスを特徴付ける形状といえるのは，ギターでは (a) のようにネックの直線の形状，鍵では (c) のように鍵溝の形状が挙げられる．(b) と (d) では鍵とギターの構造的に対応する部分の形状が抽出されている．提案手法ではこのような位置における差異を検出し，利用していることが分かる．

表 2: 各分類と正答率

	k-NN	Local Shapelet	HMM	提案手法
Control	0.880	0.930	0.953	**0.977**
CBF	0.973	0.823	**0.980**	0.977
輪郭	-	0.725	0.825	**0.850**
速度	-	0.750	**1.000**	**1.000**



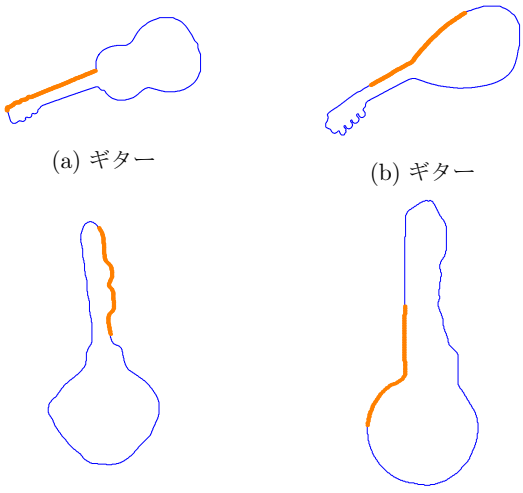
(a) 探索

(b) 探索

(c) 追跡

(d) 追跡

図 3: 速度プロファイルにおける一様クラスタの例



(a) ギター　(b) ギター

(c) 鍵　(d) 鍵

図 4: 輪郭データにおける一様クラスタの例

6. おわりに

本研究では，時系列データからより正確性の高い分類を行うため，部分形状との類似度に基づく特徴量を生成した．従来研究との比較を行った結果，安定して従来手法と同等もしくは上回る正答率を達成した．また，分類に用いられた部分形状を視覚的に抽出し，人間が判別を行う方法と類似した形で利用していることを表すことができた．

参考文献

[1] Chang, Chih-Chung and Lin, Chih-Jen, LIBSVM: A Library for Support Vector Machines, ACM Trans. Intell. Syst. Technol. pp.1-27, 2011.
[2] C.M. ビショップ, パターン認識と機械学習　(下), シュプリンガー・ジャパン, 2008
[3] Jon Hills, Jason Lines, Edgaras Barabauskas, Classification of time series by shapelet transformation, Data Mining and Knowledge Discovery, pp.1-31, 2013.
[4] K. Bache and M. Lichman, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, University of California, Irvine, School of Information and Computer Sciences, 2013.
[5] Kadous, Mohammed Waleed, Temporal Classication: Extending the Classication Paradigm to Multivariate Time Series, Doctoral Dissertation University of New South Wales, 2002
[6] Latecki, Longin Jan and Lakämper, Rolf and Eckhardt, Ulrich, Shape Descriptors for Non-rigid Shapes with a Single Closed Contour, Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.424-429, 2000.
[7] Mueen, Abdullah and Keogh, Eamonn and Young, Neal, Logical-shapelets: An Expressive Primitive for Time Series Classification, Proceedings of the 17th ACM

[*3] www.doc.gold.ac.uk/~mas02mg/software/hmmweka/index.html

SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, pp.1154-1162, 2011.
[8] Xi, X. Keogh, E. Shelton, C. Wei, L., Ratanamahatana, C. A., Fast Time Series Classification using Numerosity Reduction, ICML '06: Proceedings of the 23rd International Conference on Machine Learning, pp.1033-1040, 2006
[9] Lexiang Ye, Eamonn Keogh, Time series shapelets: a novel technique that allows accurate, interpretable and fast classification, Data Mining and Knowledge Discovery, Volume 22, Issue 1-2, pp.149-182, 2011.
[10] Z. Xing, J. Pei, P. Yu, and K. Wang., Extracting Interpretable Features for Early Classification on Time Series. In Proceedings of the 11th Siam International Conference on Data Mining (SDM11), pp.247-258, 2011.
[11] 上原邦昭, 川村俊樹, 近傍事例集合の分布密度を用いたMultiple-Instance 学習, 情報処理学会研究報告, (数理モデル化と問題解決研究報告), pp.65-68, 2007.
[12] Kouno, A.; Takano, S., Suzuki, E., Constructing Low-cost Swarm Robots that March in Column Formation, Proceedings of the 7th International Conference on Swarm Intelligence, Springer-Verlag, pp.556-557, 2010.
[13] 須賀祐太郎, 安藤晋, 関庸一, 人行動分類のための類型パターンに基づく最近傍法, 情報処理学会研究報告, (数理モデル化と問題解決研究報告), pp.1-5, 2013.
[14] 辻本貴昭, 上原邦昭, Local Shapelet を用いた時系列分類に最適な距離尺度の選択, 情報処理学会研究報告, (バイオ情報学) 2012-BIO-32(27), pp.1-6, 2012.

本論文に表記の誤りがありました。
以下の様に訂正させて頂きます。

頁	行	誤	正
3	27	X_i の部分系列集合 S_i は $\cup_{i=1}^{n}\mathbf{s}_i$ と表される.	X_i の部分系列集合は $S_i = \{\mathbf{s}_t\}_{t=1}^{T-l+1}$ と表される.
	28	全訓練データの部分系列の和集合を $\mathcal{S} = \{S_1, \cdots, S_n\}$ と記す.	全訓練データの部分系列の和集合を $\mathcal{S} = \bigcup_{i=1}^{n} S_i$ と記す.
	38	本研究では $\{\mathbf{v}(X_i)\}$ を 特徴量 として	本研究では $\{\mathbf{v}(X_i)\}$ を 入力 として
	42	$\{\mathbf{v}X_i\}$ を X_i のハウスドルフテンプレート変換 (Hausdorff Template Transform:HaTT) と呼ぶ.	$\mathbf{v}(X_i)$ を X_i のハウスドルフテンプレート変換 (Hausdorff Template Transform:HaTT) と呼ぶ.
	表 1 見出し	事例数 クラス数 分割比 時系列長	クラス数 事例数 分割比 時系列長
4	10	交差検定における パラメータ k は $k = 5$ である.	交差検定における 分割数は 5 である.