

# Shapeletsによる時系列特徴抽出と オンラインギャンブルでの問題行動の予測

Time-series Feature Extraction by Shapelets and  
Prediction of Problem Behavior in Online Gambling

中村 凌子 <sup>\*1</sup> 鈴木 浩子 <sup>\*2</sup> 渡部 勇 <sup>\*2</sup> 高木 友博 <sup>\*1</sup>  
Ryoko Nakamura Hiroko Suzuki Isamu Watanabe Tomohiro Takagi

<sup>\*1</sup>明治大学大学院理工学研究科情報科学専攻 <sup>\*2</sup>富士通研究所  
Department of Computer Science, Meiji University Fujitsu Laboratories Ltd.

In recent years, the field of behavior analysis using online gambling data has developed. However, researches on time-series behavioral changes are inadequate. In this study, we propose a classifier that quantifies the changes of the player's time series of online gambling behavioral data using distance measurement with Shapelet for the purpose of early detection of players leading to problem gambling. Especially, we investigate the prediction capabilities of local Shapelets which represent short term user behavior characteristics and global Shapelets representing long term ones. Prediction experiments show that the time series features were more effective than the non-time series features, and also show that the case using both the local time series features and the global time series features performed the best.

## 1. はじめに

近年、国内においてカジノ開業に向けた議論が進む中、ギャンブル依存症の早期発見・予防のための研究のニーズが高まっている。特に、オンラインギャンブル業者のデータを用いた行動分析の分野では、問題ギャンブリング（ギャンブルにより生活に問題が発生している状態、以下 PG と呼ぶ）に至るプレイヤーの早期発見を目的とした研究などが発展してきている。ここで、PG に至るプレイヤーの早期発見において、プレイヤーの時系列の行動変化は重要な分析対象である。しかし、プレイヤーによってキャリア期間（初めてプレイを開始してからの日数）が異なることや、訪問の頻度などによって不規則に変動する値であることから時系列の行動変化を定量化することは難しく、十分な研究がなされていない [Chagas 2017]。

そこで本研究では、時系列の行動変化を表す Shapelets を用いた PG に至るプレイヤーの予測を行う。また、先行研究でみられるような、行動特性や属性などの非時系列特徴量のみを用いた分類器の精度と比較することで、時系列特徴量の重要性を示す。さらに、局所的な時系列の行動変化と大局的な時系列の行動変化に注目し、どちらの行動変化の特徴量も分類に有効な特徴量であることを実験によって示す。

## 2. 関連研究

オンラインギャンブルのデータを用いた過去 10 年の行動分析に関する研究は [Chagas 2017] で調査されているが、非時系列特徴量を用いて分析を行っている場合が多く [Gray 2012, Braverman 2013], PG に至るプレイヤーの時系列の行動変化の分析は十分に行われていない。

分類において特徴的な時系列を抽出する方法として、複数の Shapelets (クラスに共通する時系列の特徴パターン) の中からクラス分類の損失関数を最小にする Shapelet を抽出する方法がある [Josif 2014]。本研究では、ここで提案されている Shapelet の生成方法 (tslearn のライブラリ) を用いる。また、

連絡先: 中村凌子, 明治大学大学院理工学研究科情報科学  
専攻, 〒 214-8571 神奈川県川崎市多摩区東三田 1-1-1,  
ry5983rsg@cs.meiji.ac.jp

分類を目的とした Shapelet を用いた時系列特徴量の生成方法として、Shapelet と時系列データの類似度を ED (ユークリッド距離) などで求め、求めた類似度を時系列特徴量とする研究 [J. Lines 2012] やそれを時間的依存関係のある多変量時系列の分類に応用させた研究 [Aaron 2017] などがある。また、時系列データの距離測定において、ED は時間における類似度を測定するのに、DTW (動的時間伸縮法) は形状における類似度を測定するのに有効であることで知られている [Saeed 2014]。そこで、本研究では ED と DTW の両方の特性を利用し、時系列の行動分類に有効な時系列特徴量を生成する。[Suzuki 2018] はオンラインギャンブルデータにおいて Shapelet を用いて時系列の行動変化を定量化し、生成した特徴量を用いた分類を決定木などで行い、識別能力の高い特徴量を生成した Shapelet の形を解析することで、行動変化の意味解釈を行なっている。ここでは、局所的な時系列の行動変化に注目している。しかし、PG に至る行動変化はプレイヤーによってキャリア期間や行動の時間幅も異なる。そのため、本研究では、大局的な時系列の行動変化にも注目し、大局的な時系列特徴量の生成や局所的な時系列特徴量の併用による分類を行う。

## 3. 提案手法

### 3.1 システム構成

本研究の提案システムの処理フローを図 1 に示す。まず、局所的な時系列特徴量、大局的な時系列特徴量、非時系列特徴量を生成する。生成した特徴量を説明変数として統合し、ロジスティック回帰により PG に至るかどうかの 2 値分類を行う。

### 3.2 時系列特徴量の生成方法

データセット  $D$  は、プレイヤー人数分の賭け金や賭け数などの多次元の時系列データ、行動特性や属性などの多次元の非時系列データ、正解データ (PG に至るプレイヤーは 1, PG に至らないプレイヤーは 0) を含む。ここで、プレイヤー  $n$  人のデータの集合を  $X = \{x_i, i = 1, \dots, n\}$ , 正解データ  $y_i = \{0, 1\}$  の集合を  $Y = \{y_i, i = 1, \dots, n\}$ , 長さ  $L$  の  $m$  次元の時系列データ  $mt_j$  の集合を  $MT_i = \{mt_{i,j}, j = 1, \dots, m\}$  とする。また、データの集合  $X$  を分割した訓練データとテストデータの集合を  $X_{train}, X_{test}$  とし、正解データを  $Y_{train}, Y_{test}$  とする。

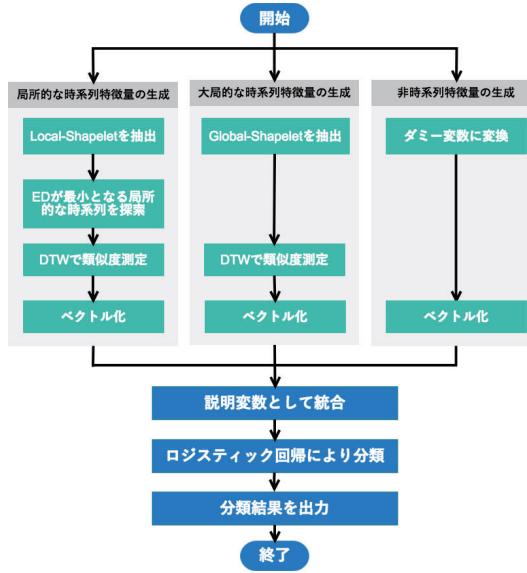


図 1: 提案システムの処理フロー

### 3.2.1 Shapelet の抽出

Shapelet は同じクラスに共通する時系列の特徴パターンであり、ここでは訓練データ  $X_{\text{train}}, Y_{\text{train}}$  から Shapelets を抽出する。さらに Shapelet と時系列データとの距離を特徴量としたクラス分類の損失関数が最小になるような Shapelets を学習する [Josif 2014] ことで、クラス分類において有効な Shapelets を抽出することができる。それを PG における分類に当てはめた場合、PG に至るプレイヤーに共通する行動を表す局所的な時系列と PG に至らないプレイヤーに共通する局所的な時系列の中から、クラス分類において有効な Shapelets を抽出することができる。このようにして、長さ  $l$  の  $k$  番目の Shapelet  $S_{k,l}$  を  $p$  個抽出した集合を Shapelets とする。

$$\text{Shapelets} = \{S_{k,l}, k = 1, \dots, p\} \quad (1)$$

### 3.2.2 局所的な時系列特徴量の生成方法

局所的な時系列変化に注目する理由は、PG に至る過程で行動変化や行動の原因となる状況の変化が部分的に現れているのではないかと考えられるためである。

まず、局所的な時系列の長さ  $l'$  ( $l'$  は時系列データの全長  $L$  より短いものとする) の Shapelet (以下、Local-Shapelet と呼ぶ) を訓練データから  $p$  個抽出する。PG に至るプレイヤーと PG に至らないプレイヤーの分類に有効な局所的な時系列変化が抽出される。

次に、抽出した Local-Shapelet に近い時系列変化がどの時間に現れているかを見つけるために、Local-Shapelet と局所的な時系列の類似度を、時間における類似度を測定するに優れている ED を用いて探索する。ここで、長さ  $L$  の時系列データ  $mt_{i,j}$  を時間区間  $[t, t + l']$  に分割した長さ  $l'$  の局所的な時系列を  $W_{t,l'}$  とし、長さ  $l'$  の Shapelet  $S_{k,l'}$  と  $W_{t,l'}$  の ED を用いた時間における類似度  $D_{\text{ED}}$  を求める。 $s_{k,l',t}$  を Local-Shapelet  $S_{k,l'}$  における時間  $t$  の値、 $w_{t,l'}$  を  $W_{t,l'}$  における時間  $t$  の値とすると、

$$D_{\text{ED}}(S_{k,l'}, W_{t,l'}) = \sum_{t=0}^{l'+1} (s_{k,l',t} - w_{t,l'})^2 \quad (2)$$

式 (2) のように定義できる。

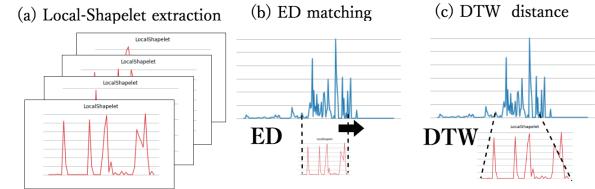


図 2: 局所的な時系列特徴量の生成方法

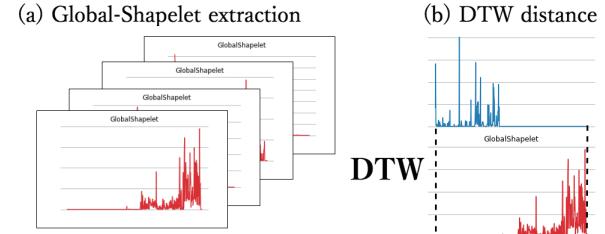


図 3: 大局的な時系列特徴量の生成方法

次に式 (3) により、 $t$  を 1 ずつ増やして (初期値は  $t = 0$  とする)、式 (2) で求めた  $D_{\text{ED}}$  をスライディングウィンドウ式に測定し、 $D_{\text{ED}}$  が最小となる局所的な時系列  $w'_{t,l'}$  を求める。

$$w'_{t,l'} = \arg \min_{w_{t,l'} \in mt_{i,j}} D_{\text{ED}}(S_{k,l'}, w_{t,l'}) \quad (3)$$

さらに  $S_{k,l'}$  と  $w'_{t,l'}$  の距離をより正確に測定するため、時間伸縮可能な DTW (動的時間伸縮法) を用い、式 (4) によって算出する。

$$D_{\text{local}}(S_{k,l'}, mt_{i,j}) = \text{DTW}(S_{k,l'}, w'_{t,l'}) \quad (4)$$

式 (4) の右辺は  $S_{k,l'}$  と局所的な時系列  $w'_{t,l'}$  の DTW 距離を表し、左辺は  $S_{k,l'}$  と  $mt_{i,j}$  の局所的類似度  $D_{\text{local}}$  を表す。この  $D_{\text{local}}(S_{k,l'}, mt_{i,j})$  の値が、 $i$  番目のプレイヤーの局所的な Shapelet に対応する特徴量となる。

### 3.2.3 大局的な時系列特徴量の生成方法

局所的な時系列特徴量のみでは、プレイヤーの一連の行動変化を定量化することはできていない。そのため、全局的な時系列特徴量を全長  $L$  の Shapelet (以下、Global-Shapelet と呼ぶ) との類似度を測定することによって生成する。

まず、全局的な長さ  $L$  の Global-Shapelet  $S_{k,L}$  を訓練データから  $p$  個抽出する。それらは PG に至るプレイヤーと PG に至らないプレイヤーの分類に有効な全局的な時系列変化を表す。

次に、Global-Shapelet  $S_{k,L}$  との距離を時間伸縮可能な DTW 距離で測定することによって、 $S_{k,L}$  との類似度を測定する。

$$D_{\text{global}}(S_{k,L}, mt_{i,j}) = \text{DTW}(S_{k,L}, mt_{i,j}) \quad (5)$$

式 (5) の右辺は  $S_{k,L}$  と  $mt_{i,j}$  の DTW 距離を表し、左辺は  $S_{k,L}$  と  $mt_{i,j}$  の全局的類似度  $D_{\text{global}}$  を表す。この  $D_{\text{global}}(S_{k,L}, mt_{i,j})$  の値が、 $i$  番目のプレイヤーの全局的な Shapelet に対応する特徴量となる。

## 4. データ

本研究で用いるデータセットは、[Gray 2012] において用いられているものと同一である。このデータセットは、2008 年～

表 1: 実験で用いた特徴量の大分類

		集計単位	集計方法	特徴量	数
時系列特徴量	局所的時系列特徴量	合計値	週次	賭け金の Local-Shapelet から生成	4
				負け金の Local-Shapelet から生成	4
		累積値		負け金の Local-Shapelet から生成	4
				賭け数の Local-Shapelet から生成	4
	大局的時系列特徴量	合計値		賭け金の Global-Shapelet から生成	4
				負け金の Global-Shapelet から生成	4
		累積値		負け金の Global-Shapelet から生成	4
				賭け数の Global-Shapelet から生成	4
非時系列特徴量	行動特性特徴量	日次		平均賭け金	1
				平均賭け数	1
				賭け金の変動係数	1
				賭け数の変動係数	1
	属性特徴量			性別	1
				年代	1

2009 年に PG に至ったプレイヤー 2,068 人と PG に至らなかつたプレイヤー 2,066 人の 4,134 人の情報を含み, RawDataset1 では 1999 年 9 月 17 日～2009 年 11 月 27 日の期間にオンラインギャンブルに加入了したプレイヤーの属性情報, RawDataset2 では 2000 年 5 月 1 日～2010 年 11 月 10 日にアクティブな行動をしたプレイヤーの日次行動情報, RawDataset3 ではプレイヤーの PG に至った情報を含む。

本研究では、このデータセット<sup>\*1</sup>から欠損値を含む 21 人を除外した 4,113 人 (PG に至ったプレイヤー 2,068 人, PG に至らなかつたプレイヤー 2,045 人) のデータを使用し, RawDataset1 から性別, 誕生日, PG の項目を, RawDataset2 から訪問数, 賭け金, 負け金, 賭け数の項目を用いた。また、日次行動情報は週次に集約したのち、アクティブな期間の長さの短いプレイヤーの空白期間に 0 を詰めることによって最も長いプレイヤーの長さ ( $L=548$ (週)) に揃えた。非時系列の行動特性特徴量は、キャリア期間全体での 1 日の平均賭け金, 平均賭け数, 賭け金の変動係数, 賭け数の変動係数などの集計値である。(表 1)

## 5. 実験

生成した特徴量を用いて分類器によってプレイヤーが将来 PG に至るかどうかの予測を行う。訓練データとテストデータを 7:3 に分割し、ロジスティック回帰で分類を 100 回行った。評価は Accuracy, Precision, Recall, F1score を用いた。

### 5.1 Shapelet の抽出

事前実験により、週次の賭け金の合計値、負け金の合計値、負け金の累積値、賭け数の累積値の 4 つの時系列変量を選択した。

#### 5.1.1 Local-Shapelet の抽出

4 つの時系列変量それぞれに対し、長さ 26 (週) の Shapelet を 4 つずつ、合計 16 個抽出し、Local-Shapelets とした。抽出された Local-Shapelets の中から 2 つの例を図 4 に示す。週次の賭け金の合計値による Shapelet(左) は、1 回少し賭けた後、3 回ほど賭け金を減らしつつ賭け、最後に最も大きく賭けるような行動変化と解釈できる。また、週次の負け金の合計値による Shapelet(右) は、1 度だけ大勝ちする様子を表している。

#### 5.1.2 Global-Shapelet の抽出

4 つの時系列変量それぞれに対し、長さ 548 (週) の Shapelet を 4 つずつ、合計 16 個抽出し、Global-Shapelets とした。抽

\*1 Division on Addiction, Behavioral characteristics of Internet gamblers who trigger corporate responsible gambling interventions. Medford, MA: Division on Addiction, The Transparency Project [database distributor], February 7, 2016.

出された Global-Shapelets の中から 2 つの例を図 5 に示す。週次の賭け数の累積値による Shapelet(左) は、全体の半分くらいまで賭け数の傾きが大きく、徐々に賭け数を減らしていつている行動変化と解釈できる。週次の負け金の合計値による Shapelet(右) は、全体を通して負け金が大きくなつても賭け続けた時期があると解釈できる。

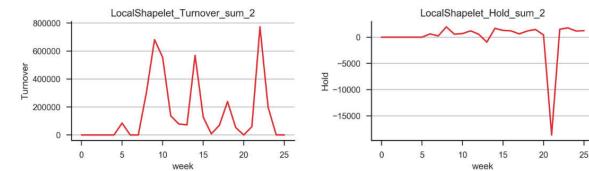


図 4: 抽出された Local-Shapelet の例

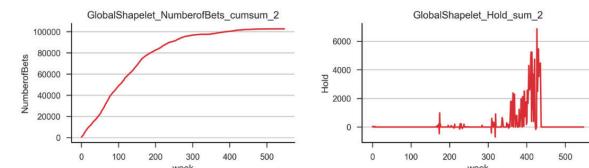


図 5: 抽出された Global-Shapelet の例

### 5.2 時系列特徴量と非時系列特徴量の比較実験

#### 5.2.1 実験方法

ここでは、生成した時系列特徴量が非時系列特徴量と比べて分類に有効な特徴量であるかどうかを調べる。6 次元の非時系列特徴量 (Non-TS で示す), 32 次元の時系列特徴量 (TS で示す), 非時系列特徴量と時系列特徴量を合わせた 38 次元の特徴量 (Non-TS+TS で示す)(表 1) による推定精度の比較をする。

#### 5.2.2 実験結果と考察

図 6 に訓練データとテストデータの分割を変えた 2 回の実験の Accuracy を箱ひげ図で示す。左図が 1 回目、右図が 2 回目の実験結果を示す。Non-TS と TS の Accuracy を比較すると、2 回とも TS による分類の方が精度が高い。全実験における各特徴量での Accuracy, Precision, Recall, F1score の平均値を表 2 に示す。表 2 で Non-TS, TS を比較すると、全ての評価値において TS の方が高い値であることが分かる。そのため、時系列特徴量の方が非時系列特徴量に比べて分類に効果があるこ

表 2: 実行結果の平均値

Feature	Accuracy	Precision	Recall	F1score
Non-TS	0.720	0.756	0.656	0.702
TS	<b>0.799</b>	0.845	<b>0.738</b>	0.787
Non-TS + TS	<b>0.799</b>	0.845	<b>0.738</b>	<b>0.788</b>
Non-TS + Local-TS	0.781	<b>0.851</b>	0.685	0.759
Non-TS + Global-TS	0.783	0.849	0.693	0.763

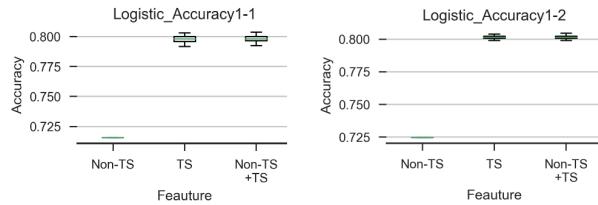


図 6: 時系列特徴量と非時系列特徴量の比較

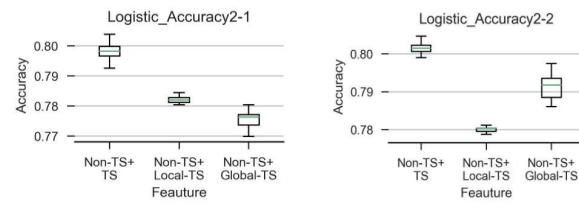


図 7: 局所的特徴量と全局的特徴量の比較

とがわかる。一方で、図 6 において、TS と Non-TS+TS の精度があまり変わらないことから、時系列特徴量が分類において大きな役割を果たしていることもわかる。

### 5.3 局所的時系列特徴量と全局的時系列特徴量の比較

#### 5.3.1 実験方法

次に、局所的時系列特徴量と全局的時系列特徴量についてどちらがまたは両方の特徴量が有効かどうか調べる。非時系列特徴量と時系列特徴量を合わせた 38 次元の特徴量 (Non-TS+TS で示す)、非時系列特徴量に局所的時系列特徴量のみ加えた 22 次元の特徴量 (Non-TS+Local-TS で示す)、非時系列特徴量に全局的時系列特徴量のみ加えた 22 次元の特徴量 (Non-TS+Global-TS で示す) による推定精度の比較をする。

#### 5.3.2 実験結果と考察

図 7 に訓練データとテストデータの分割を変えた 2 回の実験の Accuracy を箱ひげ図で示す。左図が 1 回目、右図が 2 回目の実験結果を示す。また、全実験における各特徴量での Accuracy, Precision, Recall, F1score の平均値を表 2 に示す。図 7 をみると、Non-TS+TS の精度が最も高いことがわかる。表 2 においても、Accuracy, Recall, F1score の 3 つの評価値において Non-TS+TS が最も高く、局所的時系列特徴量と全局的時系列特徴量の両方を含めた方が分類に有効であることが分かる。

また、Non-TS+Local-TS と Non-TS+Global-TS を比較すると、図 7 から 1 回目 (左図) では Non-TS+Local-TS の方が精度が高く、2 回目 (右図) では Non-TS+Global-TS の方が精度が高いことが分かる。表 2 では、Precision のみ Non-TS+Local-TS が最も高いが、Accuracy, Recall, F1score の 3 つの評価値において Non-TS+Global-TS の方が高い。これから、局所的時系列特徴量と全局的時系列特徴量の優劣をつけることはできないと考えられる。

## 6. おわりに

本研究では、オンラインギャンブルデータを用いて、時系列の行動変化を Shapelet を用いて量化し、局所的な時系列特徴量と全局的な時系列特徴量を生成した。これらの特徴量を用いた問題ギャンブリングに至るかどうかの予測実験から、時系列特徴量が非時系列特徴量より行動予測に効果があること、局所的な時系列特徴量と全局的な時系列特徴量はどちらも有効な特徴量であることなどが分かった。今後は Shapelet 以外の時系列特徴量の生成手法の有効性についても検証したい。また、本提案手法をマーケティングの顧客行動予測などの分野に応用することも検討したい。

## 参考文献

- [Chagas 2017] Bernardo T. Chagas, Jorge F. S. Gomes, Internet Gambling: A Critical Review of Behavioral Tracking Research, Journal of Gambling Issues No.36(2017)
- [Gray 2012] Heather M. Gray, Debi A. LaPlante, and Howard J. Shaffer.: Behavioral Characteristics of Internet Gamblers Who Trigger Corporate Responsible Gambling Interventions, Psychology of Addictive Behaviors, 2012, Vol. 26, No. 3, 527535 (2012)
- [Braverman 2013] Braverman, J., LaPlante, D. A., Nelson, S. E., Shaffer, H. J.: Using Cross-game Behavioral Markers for Early Identification of High-risk Internet Gamblers, Psychology of Addictive Behaviors(2013)
- [Josif 2014] Josif Grabocka, Nicolas Schilling, Martin Wisztuba, Lars Schmidt-Thieme: Learning Time-Series Shapelets, KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (2014)
- [J. Lines 2012] J. Lines, L. Davis, J. Hills, and A. Bagnull: A shapelet transform for time series classification, KDD '12 Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2012)
- [Aaron 2017] Aaron Bostrom, Anthony Bagnall, ArXiv e-prints: A Shapelet Transform for Multivariate Time Series Classification (2017)
- [Saeed 2014] Saeed Aghabozorgi, Teh Ying Wah, Tutut Herawan, Hamid A. Jalab, Mohammad Amin Shaygan, and Alireza Jalali: A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique, The Scientific World Journal Volume 2014(2014)
- [Suzuki 2018] Hiroko Suzuki, Ryoko Nakamura, Tomohiro Takagi: Time Series Analysis on Behavioral Changes Leading to Problem Gambling, NAGS 28th Annual Conference(2018)