様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １６ 日現在

機関番号：２５４０６
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１５５０

研究課題名（和文）対話型機械学習に基づく複合イベント処理ルールの生成

研究課題名（英文）Complex Event Processing Rule generation based on Interactive Machine Learning

研究代表者
岡部　正幸（Okabe, Masayuki）

県立広島大学・地域創生学部・准教授

研究者番号：５０３６２３３０

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では，複合イベント処理におけるイベント－アクションルールの生成を対話型機械学習に基づいて高品質かつ効率的に行う方法について提案した．提案手法は主に次の2点からなる．1つ目は高性能なイベント検知方法の提案である．この研究ではShapeletと呼ばれる特徴的な時間変動に基づくデータ分類方法をベースに，Shapelet間の順序情報を追加することにより精度を向上させる方法を提案した．2つ目は効率的なルール選択方法の提案である．この研究では，同等の品質をもつ複数のルールがあった場合に，精度と妥当性の高いルールをランキング形式でユーザに提示することにより対話的に効率よく選択する方法を提案した．

研究成果の学術的意義や社会的意義
本研究の成果におけるShapeletの順序情報を考慮したストリームデータ分類方法は判定モデルがどのような時間変動に基づいて分類を行っているのか解釈しやすいという利点を維持しつつ，精度を向上させている点が従来にはなく学術的意義が高い．また，提案方法ではルール形式でモデルを生成できるため，複合イベント処理システムにおけるイベント－アクションルールとしての実装も容易になると考えられる．一方，効率的なルール選択方法の提案については，従来型システムでは多大な時間と労力を必要とした作業を機械学習の利用により軽減できる可能性があり，実システムにおける利用が期待される．

研究成果の概要（英文）：In this study, we proposed a high-quality and efficient method for event-action rule generation in complex event processing based on interactive machine learning. The proposed method consists of two main parts. The first one is to propose a high-performance event detection method, which is based on a stream data classification method using characteristic time-varying subsequences called shapelets. The second one is to propose a method for efficient interactive rule selection by presenting the rules with the highest accuracy and validity in a ranking format to the user when there are multiple rules of equal quality.

研究分野：知能情報学

キーワード：ストリームデータ分類　対話型機械学習

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

複合イベント処理（Complex Event Processing, CEP）は，実世界から時々刻々と発生するストリームデータを対象に，特定のデータ系列の発生に応じて処理を実行する技術であり，ユーザの状況に応じたタイムリーなサービス提供手段として活用できる．CEP は，金融業，小売業，製造業，流通業をはじめとする幅広い分野において，ビジネスチャンスを創出する新たなアプローチとして期待されている．CEP の実行は，ストリームデータにおけるイベントと呼ばれる特定のデータ系列の発生を条件とし，それらに適切なアクションを対応付けたルールに基づいて行われる．例えば，株取引に用いた場合は，株価の特定の変動をイベントとし，各イベントに応じて売買注文というアクションを実行することができる．また，マーケティングに用いる場合は，特定のプロファイル情報（年齢，性別など）をもつユーザが特定のエリアに居る状態をイベントとし，そのユーザに向けて限定クーポンを発行するといったことができる．このように，CEP は複数のイベントとアクションからなるルールに基づいて実行されるため，その効果は定義されるルールの品質に依存する．現在，CEP のルール生成は，人間がプログラムとして書き下しており，本格的な CEP 運用を行うためのルール集合を準備するまでに通常数ヶ月を要すると言われている．また，運用開始後も継続的にルールの改修を行う必要があり，運用開始の遅れによるビジネスチャンスの喪失や企業規模によっては導入ハードルが高いといった問題がある．

２．研究の目的

本研究の目的は，先に述べた CEP 導入時の問題点を解決するため，CEP の要であるイベント–アクションルールの生成を，対話型機械学習を用いて高品質かつ効率的に行う方法を提案することである．具体的には，以下の 2 点について検討する．

I. 高性能なイベント検知方法の提案
II. 効率的なルール選択方法の提案

Ⅰについては，ルール内のアクションはストリームデータ内におけるイベントの発生に基づいて実行されるため，イベント検知の精度はルールの品質に大きな影響を与える．よって，高性能な検知アルゴリズムの開発が必要となる．ただし，誤ってアクションを実行してしまうことがないよう，アルゴリズムがどのような根拠に基づいて検知を行っているかについて可視化する必要がある．Ⅱについては，同等な性能をもつ複数のルールが生成される場合や，判定根拠の妥当性が不明なルールが生成される場合に，それらの中から適切なルールの選択を効率よく行う方法を提案することである．一般に機械学習ではハイパーパラメータを変える，訓練データ集合を変えるなどの操作を行うことにより様々なモデルを生成することができるが，本研究では，性能面だけでなくルールの妥当性も考慮したモデル選択を行う方法について検討する．

３．研究の方法

本研究における 2 つの目的について，以下のように研究を行う．

I. 高性能なイベント検知方法の提案

本研究では，イベント検知をイベント発生の有無を正解ラベルとするストリームデータ分類問題として捉え，機械学習に基づく検知アルゴリズムの開発を行う．このアルゴリズムでは，モデルの判断根拠を可視化する必要があるため，Shapelet に基づく分類方法を採用する．Shapelet はデータ系列中における特徴的な変移をもつ部分系列であり，これを特徴として利用することにより，分類時に重要な役割を果たす Shapelet の抽出が可能となる．また，Shapelet が示す変移は可視化が容易であり，イベント検知の判定根拠として利用することも可能である．本研究では，更に検知性能を向上させるため，Shapelet 間の順序情報の導入を試みる．これにより，ルールにおけるアクションの実行条件としてイベント発生の有無だけでなく，その発生順序も指定できることになり検知性能の向上が期待できる．

II. 効率的なルール選択方法の提案

Ⅰのアルゴリズムを用いて検知されたイベント集合とアクションを紐付け，訓練データとすることで決定木などの機械学習アルゴリズムを用いてルールの生成が可能である．ただし，生成できるルールについて，実際の運用前に致命的な誤判定を起こさないかチェックする必要がある．また，生成可能なルールは１つとは限らず，同等の性能をもつ複数のルールが生成可能な場合には，それらの中から適したものを選択する必要がある．このような検証プロセスを実現するため，本研究では，決定木モデルを対象に，生成された複数のモデルの中から判定根拠の妥当性と性能のバランスの取れた良いモデルを効率よく探索する方法を提案する．提案方法では，選択候補となる決定木モデルそれぞれについてモデルを特徴づける指標を設定し，バランスの取れたモデルを自動的にランキングする．

４．研究成果

I. 高性能なイベント検知方法の提案について

ストリームデータ分類は，状態推定や行動認識などの応用をもつデータ解析技術の一つであり，その実現方法として様々なものが提案されている．その中でも Shapelet と呼ばれる特徴的な部分系列集合に基づく分類手法は，高性能な手法の一つとして知られている．Shapelet を用いた分類手法では，これまで個々の Shapelet は独立した特徴として扱われ，その出現順序については考慮されてこなかった．しかし，同じ種類の Shapelet が出現するデータであっても，その出現順序が異なれば，異なるクラスに属する時系列データであるとした方が良い場合もあると考えられる．そのため，本研究では Shapelet 間の相対的な順序関係も分類時の特徴として利用する方法を提案し，その有効性を実験により検証した．

Shapelet に基づくデータ分類では，特徴的な部分系列を Shapelet 集合として抽出し，各データを個々の Shapelet との最小距離を要素としてもつ特徴ベクトルに変換することで分類が行われる．本研究では，Shapelet 集合の抽出方法として，Grabocka らが提案したランダムサンプリングおよび最近傍法による選択に基づく方法を用い，抽出された Shapelet 集合について，各ストリームデータとの最小距離を計算するだけでなく，データ中における Shapelet の出現順序情報も取得する．この出現順序情報の取得は文字列カーネルを利用して行われる．文字列カーネルは，2 つの系列データそれぞれについて，生成しうる長さ n の全文字列のうちどれが実際に出現するかどうかを調べ，出現した文字列の一致度を計算したものである．つまり，文字列カーネルの各要素は 2 つの系列データにおいて出現する文字の順序がどれくらい似ているかを示している．本研究では，系列データにおける文字を Shapelet に置き換えることで，2 つの時系列データにおける Shapelet の出現順序がどれくらい似ているかを数量化する．

提案手法では，従来手法における Shapelet との最小距離情報と Shapelet 間の順序情報のどちらも利用するため，それぞれをカーネルで表現し，重みを用いて線形結合する．前者を最小距離カーネルK^{md}，後者について，特に長さ n の文字列カーネル（n-gram カーネルと呼ばれる）をK_i^{ng}としたとき，それらを統合したカーネルK^{mul}は次のように表すことができる．

$$K^{mul} = \mu_1 K^{md} + \sum_{i=2}^{n} \mu_i K_i^{ng}$$

ここで，n-gram カーネルについては，n の値を変化させた複数のカーネルを用いることを想定している．複数のカーネルに重みを設定することで，データに応じて適切な値を自動的に設定できるようにした．式中の　μ_i が各カーネルの重みを表しており，カーネルアライメントによるマルチカーネル学習をベースとした方法を用いてその値を求めている．

提案手法の性能を 4 つのベンチマークデータを用いて評価した．実験では，K^{mul}を用いた提案手法とK^{md}のみを用いた従来方法とを比較し，n-gram カーネルを用いた場合の有効性について調べた．Shapelet 抽出時の繰り返し回数は 10 万回，Shapelet の長さはデータ長の 1/5，2/5，

3/5 とした．また，分類モデルの生成は SVM を用いて行い，2 つの手法では同一の Shapelet 集合を用いた．実験結果を表 1 に示す．表中の「なし」は従来手法，その他の「ng = #」は提案手法を表している．実験結果より，いずれのデータにおいても，提案手法の精度が従来手法を上回っており，その効果を確認することができた．

表 1　実験結果

データ	ngram	精度	±	標準偏差	データ	ngram	精度	±	標準偏差
ECG200	なし	0.8540	±	0.0312	GunPoint	なし	0.9548	±	0.0310
	ng=2	0.8764	±	0.0234		ng=2	0.9628	±	0.0225
	ng=2,3	**0.8776**	±	0.0228		ng=2,3	0.9660	±	0.0213
	ng=2,3,4	0.8752	±	0.0235		ng=2,3,4	0.9672	±	0.0207
	ng=2,3,4,5	0.8744	±	0.0241		ng=2,3,4,5	**0.9673**	±	0.0201
Medical Images	なし	0.6889	±	0.0139	Yoga	なし	0.6511	±	0.0272
	ng=2	0.7225	±	0.0121		ng=2	0.7147	±	0.0196
	ng=2,3	**0.7238**	±	0.0107		ng=2,3	**0.7164**	±	0.0225
	ng=2,3,4	0.7237	±	0.0108		ng=2,3,4	0.7156	±	0.0224
	ng=2,3,4,5	0.7237	±	0.0108		ng=2,3,4,5	0.7158	±	0.0206

II. 効率的なルール選択方法の提案

本研究では，説明可能な AI（eXplainable AI，XAI)の手法を用いて，複数生成された判定モデルの中から精度だけでなくモデルの解釈可能性も考慮した選択方法を提案する．解釈可能性とは機械学習モデルが予測, 推定結果に至ったプロセスを, 人間が解釈可能であるかどうかを指す．従来研究の多くでは，この解釈可能性をあらかじめ人間が定義した尺度に置き換えて最適化するという方法がとられていたが，モデルの評価を人間が直接行なった方が，解釈可能性に関する人間の意向をよりモデルに反映させやすい．特に CEP におけるイベントアクションルールの生成においては，イベント検知の妥当性は尺度に置き換えることが難しく，人手による検証が不可欠となる．解釈可能性を考慮する場合，モデルの評価には人間の介在が必要となるが，全てのモデルを人間に評価させることは難しい．本研究では少ない評価回数で解釈性の高いモデルを導き出せるように人間に評価させるモデルを選択する方法を提案する．

本研究では決定木モデルを対象にしたモデル選択方法を考える．一般に決定木は木の深さを増やすことでモデルの精度を向上させることができるが，同時にモデルの判定根拠の妥当性が低下し，過学習を引き起こすことにつながる．このため，精度と妥当性の 2 つの観点においてバランスの取れたモデルを選択することが重要となる．選択候補となる決定木は，ハイパーパラメータを変化させることによって生成した．本研究では，モデルに与える影響度が高いと考えられるハイパーパラメータ 3 つを選択した．それぞれ，決定木の深さの最大値，木の成長における早期停止の閾値，葉に属する最小サンプル数を設定するものである．これら以外のハイパーパラメータについてはデフォルト値を用いた．これら 3 つのパラメータについて，それぞれ設定値を 10，10，5 パターン用意し，合計 500 パターンのモデル候補の中から選択を行うこととした．モデルの選択手順を以下に示す．

1. 構築した 500 パターンのモデルからランダムに 5 つのモデルをユーザに提示する．
2. ユーザに 5 つの中からポジティブだと思うモデルを複数選択してもらう．
3. その選択をもとに正解ラベルを作成しランキング学習を行う．
4. 学習して取得したランキングスコアが高い上位 5 つのモデルをユーザに再度提示する．
5. 1～4 を適切なモデル選択ができるまで繰り返す．

ユーザがモデルを評価する時の指標は，システムから人間に提示されるモデルの説明である．モデルの説明は一般的には重要度の高い特徴などを人間に示すことにより行われる．本研究では決定木モデルを説明するために，まず決定木内に含まれる判定ルールを抽出し提示した．こ

れは決定木そのものを提示するよりも個別の判定ルールを閲覧できるようにした方がより解釈性が高まると考えたからである．次に特徴の重要度を提示した．重要度の算出方法には Permutation Importance(以下 perm)と SHAP Value(以下 SHAP)を利用した．これら判定ルールと特徴の重要度に関して更に細かい指標を提示した．それらの指標を以下に示す．

- ルールの個数
- 各ルールが被覆するデータ数の平均値と標準偏差
- 各ルールのクラスの偏りの平均値と標準偏差
- Perm/SHAP が 0 より大きい値を持つ特徴の個数
- Perm/SHAP の平均値と標準偏差
- テストデータに対する予測精度

これら 9 つの指標はユーザの判断材料としてではなく，主にランキング関数の学習材料として用いる．ランキング関数の学習にはロジスティック回帰を用いた．学習は，モデル選択手順においてユーザが選択したモデルをポジティブ，それ以外の選択されなかったモデルをネガティブとして 2 クラスの分類が行われ，対応する分類器が生成される．この分類器をユーザがまだ評価していないモデルに適用することにより，各モデルの評価値が計算される．この評価値をもとにランキングを行い，上位 5 つをユーザに提示する．

提案手法を国勢調査データの分析に適用し，提案手法の有効性を検証した．国勢調査データは教育年数や職業，人種，性別，配偶者の有無などの説明変数から年収が 50,000 ドルを超えるかどうかを予測することを目的としている．今回はその分類に至った判断根拠を示すモデルをユーザに評価してもらう．図 1 にユーザに評価してもらうモデルの表示内容を示す．提案手法の有効性は，学習なしのモデル提示を行い比較検討した．提案手法の学習ループを 3 回と設定し，初期検索の 5 個と合わせて計 20 個のモデルを評価する．学習なしのモデル提示でも同様に 20 個のモデルに対して評価を行う．提案手法の有効性は，このポジティブ評価されたモデルの個数を比較することで測る．今回の実験は予備実験として参加者 1 名により行った．実験では，モデルをランダムに提示する方法と提案手法によるランキングに基づいて提示する方法の比較，重要説明変数の算出方法として perm と SHAP の 2 つの比較，更にモデルの判定基準として簡潔なモデルを望ましい場合と複雑なモデルが望ましい場合を想定し，それぞれ意図通りのモデルがランキング上位に提示されるかについて調べた．まず，モデルの提示方法については，提案手法に基づく方法の方が，ポジティブ評価だったモデル数が平均 5 個以上と大きく上回った．重要説明変数の算出方法の比較については，本実験においてはポジティブ評価数に大きな差はないものの全体的に SHAP よりも perm の方が有効であった．最後に，ランキング上位に提示されるモデルが予め設定した判定基準に追従しているかどうかについては，概ねユーザが選択した特徴を持つモデルが上位に提示されるようになることが確認できた．以上のことから，提案手法の有効性が確認できたとともに，モデルのランキングを行う際に利用したルールの個数などの指標が有効に機能していたことが分かった．

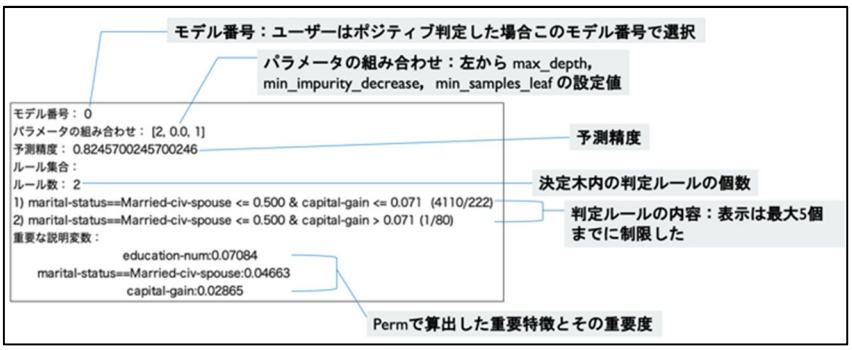

図 1　ユーザに提示するモデルの説明

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 折中日花莉，岡部正幸
２．発表標題 適合フィードバックによる解釈性に優れた予測モデルの探索
３．学会等名 2020年度（第71回）電気・情報関連学会中国支部連合大会
４．発表年 2020年

１．発表者名 藤岡公平，岡部正幸
２．発表標題 Shapeletの順序性を考慮した時系列データ分類
３．学会等名 2020年度（第71回）電気・情報関連学会中国支部連合大会
４．発表年 2020年

１．発表者名 藤岡公平，岡部正幸
２．発表標題 Shapeletの順序性を利用した時系列データ分類
３．学会等名 2021年電子情報通信学会総合大会
４．発表年 2021年

１．発表者名 岡部 正幸
２．発表標題 アンサンブル学習に基づく半教師ありクラスタリング
３．学会等名 電気学会 スマートシステムと制御技術シンポジウム 2019
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関