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　２ 年　 ５ 月 １６ 日現在

機関番号：１３３０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００００９６

研究課題名（和文）選択的に実行履歴を記録する手法の改善と新しい応用の開発

研究課題名（英文）Development of an improved method for selective recording of execution history and its new applications

研究代表者
櫻井　孝平（Sakurai, Kohei）

金沢大学・電子情報通信学系・助教

研究者番号：８０５９７０２１

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：本研究は大規模なデータに対する処理方法に対して検討を行った結果，木構造のモデルを使ったデータ処理をアクターモデルを使った並列分散環境上で実現する手法を提案・開発した．本研究の手法では，木構造のノードをアクターとする設計のパターンにより，オンラインの分類木と階層型クラスタリングのような異なるモデルに対応し，高速で大規模なデータの処理を可能にするシステムが実現できることを実験により示した．

研究成果の学術的意義や社会的意義
本研究の成果は既存の機械学習アルゴリズムを大規模なデータにシームレスに対応させるための手法を提案している．提案手法によりデータの分類やクラスタリングなどを扱うシステム開発が，多くの開発者が慣れ親しんだ手法により理解しやすいモデルの定義によって迅速に行うことが可能となる．結果としてデータの分析に関する多くの変更や性能の向上に関する要求に対応が容易となる．

研究成果の概要（英文）：In this study, we considered methods for large scale data processing, and we proposed and developed a method for data processing in tree models with the parallel and distributed environment using the actor model. Our method can cope with multiple models including online classification trees and hierarchical clustering with a design pattern which describes tree nodes as actors, and we showed that it can effiicently handle actula large scale data inputs from our experiments.

研究分野：プログラミング

キーワード：大規模データ処理　アクターモデル　機械学習

１．研究開始当初の背景

(1) プログラムの実行履歴の記録技術は逆戻りデバッグのようなソフトウェア開発支援のために研究・開発されてきた．これはプログラムの実行時に実行した箇所や実行時のデータを実行履歴として記録する技術で，プログラムに記録のための追加的なコードを挿入する．実行履歴は永続的なストレージに記録され，後に再生することでデバッグ等に利用するが，プログラムの実行で生成される値であり巨大なデータとなる．

(2) そのような巨大なデータを効率よく解析し役立てる方法が求められている．特にビッグデータと呼ばれるような大規模なデータとして，それらを機械学習の手法を適用するような需要が高まっている．

２．研究の目的

(1) 本研究の当初の目的として，プログラムの実行履歴の選択的な記録技術の改善があった．実行履歴の記録を，スケーラビリティを保ったまま，よりデバッグに役立つ情報とするため，欠陥により関係しやすい箇所 を選択し，限られた実行履歴の情報から詳細で正確な情報を抽出する技術を開発する必要があったためである．

(2) また，既存の実行履歴の用途が限定されているという問題があり，大掛かりな機構が必要な一方で，デバッグに利用するだけでは，開発現場での導入の動機付けが弱い．実行履歴はプログラムの一般的な情報であるので，選択的に記録した実行履歴をデバッグ以外に利用することも期待される．それは開発後の支援だけでなく，アプリケーションのプログラミングそのもののために，開発者が扱いやすい API を通じて実行履歴の情報を活用することも含む．

３．研究の方法

(1) 本研究では実行履歴のような大規模なデータを扱うプログラムの効率の良い開発・実装のための手法を提案・開発する．そのような目的のためには近年注目されている既存の手法をまず調査・検討し，候補となる技術を選定する．

(2) 大規模なデータを扱うアプリケーションとして需要のある機械学習を想定する．実際に公開されたデータに基づいてアプリケーションソフトウェアを開発しながら，既存の技術を適用し，実験により評価し，改善となる手法を提案する．

(3) 提案する手法も同様に評価のための実装を開発し，実際のアプリケーションソフトウェアをその上で実装した上で，実験を行なって評価する．

４．研究成果

(1) 機械学習を含むデータ処理アプリケーションでは大規模データの取り扱いと開発の迅速さが重要である．既存の研究によると，機械学習分野のエンジニアは，モデル開発の経験において，モデルのスケーラビリティや変更への対応に苦労する傾向が明らかになっている．

(2) そのための技術として，機械学習アプリケーションのための漸進的なオンラインアルゴリズムが数多く開発されている．これらのアルゴリズムを用いることで，メモリに直接格納することなく，大量のデータを次々と処理し，学習したモデルを必要に応じて迅速に取得することが可能となる．

(3) オンライン機械学習アルゴリズムは豊富な既存研究があるが，本研究では，その中でも特に，決定木や階層的クラスタリングを含む，木の学習モデルを構築する機械学習に注目した．これらの木に基づくモデルのアプリケーションもまたよく研究されており，その中には大規模データに着目し，漸進的なオンラインアルゴリズムの形態をとるものもある．このようなアルゴリズムは，入力データを軽量な表現として要約し，それをモデルとしてメモリ上に保持したまま動的にツリーノードを構築することができる．

(4) そのようなオンラインの漸進的なアルゴリズムで扱うモデルは一般的に元のデータに比べて小さいサイズとなるが，近年現実となっているような大規模なデータに対しては，既存の手法でも扱いが困難になる程大きいサイズとなり，単一のCPUによるシングルスレッドの計算や，単一の計算機のメモリ空間では効率よく扱うことが難しくなっている．本研究では並列処理・分散コンピューティングをサポートした統一的な開発方法のもとで，そのような大規模データを学習するシステムを構築する手法が確立されていないことに注目した．

(5) 本研究では，並列・分散処理を実現するための開発手法として，アクターモデルを採用する．アクターモデルは，アクターと呼ばれる計算の単位を利用し，アクター間でデータを伴う非同期

のメッセージを送信し合うことで計算が進むプログラミングの手法である．アクターは状態を持つ並列・分散処理の単位として，機械学習アルゴリズムのモデルを表現することができる．アクターモデルを利用する大きな利点として，従来の手続き型のアプリケーションの拡張として，並行メッセージが導入されることで，技術者が容易に理解し適用できることにある．その結果，コードを大幅に書き換えることなく，自然に分散コンピューティングをサポートすることが可能なり，時間的には並列処理，空間的には分散処理によるスケーラビリティを実現することが可能となる．

(6) 本研究では，木構造のモデルの構築アルゴリズムを2種類に分類し，それらにアクターモデルを適用する場合の統一的な開発方法を確立した．これは，木のサブツリーノードをアクターと定義し，木へのデータ点の入力をアクタ間のメッセージの流れとする設計パターンとして定義できる.本研究ではこのパターンをActor Tree Proxy Patternと呼んでいる．

(7) 本研究では，既存のオブジェクト指向言語のアクターモデルライブラリを用いて，定義したパターンを実装するシステム開発を行った （図1:概要)．このアプローチでは，木の中のノードをアクタとして定義し，各入力データをノードへのメッセージとして定義する．データはアクタのツリーのルートからメッセージとして入力され，ルートは木の子ノードに対して，アクタ間で同時並行的にメッセージを送信することでデータを末端の葉ノードまで届ける．アクターのノードは学習アルゴリズムに対応した内部状態と手続きをもち，データの入力に従って子ノードを増加させ，木を学習モデルとして成長させていく．

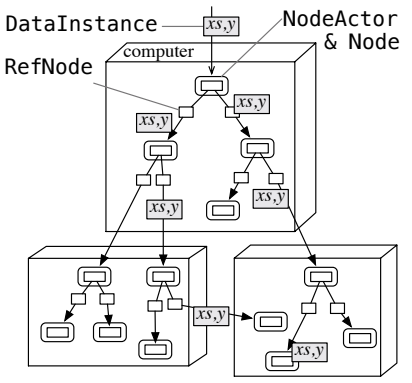

図1:アクターを使ったシステムの概要

(7) 本研究では，機械学習アプリケーションの具体的な適用例として，決定木と階層的クラスタリングの2つを取り上げた．

① 決定木としてのVFDT: 決定木はデータの分類に利用され，VFDTは木のノードが入力されたデータの統計的な特徴を保持しながら学習・分類を進めるアルゴリズムである．木がルートから末端に向けて成長するトップダウンの構築方法で，この種のアルゴリズムはルートが常に固定されることで，大規模なデータの入力がルートに集中してボトルネックになることで，性能が向上しないことが問題となる．本研究ではルートの複製を作り，定期的に一貫性を保ちつつそれらの状態を同期する機構を提案して問題を解決した．

② 階層的クラスタリングとしてのBIRCH: 階層的クラスタリングはその名の通りデータの階層的なグループ化に利用され，BIRCHは木のノードがグループの特徴の要約を保持し，それらによってバランス木の一種を構築して学習・分類を進めるアルゴリズムである．木が末端からルートに向けて成長するボトムアップのアルゴリズムであり，この種のアルゴリズムでは並行に成長する際のノード間の整合性が問題となる．本研究では横方向の追加的なノード間のリンクを導入することで，整合性を維持する手法を導入し,問題を解決した．

(8) 本研究では，アクターモデルの実用的なライブラリであるAkka Java上に提案手法のシステムを実装し，実際の公開されたデータセットを用いたシステム上での学習タスクの検討を行った．実験では，より実践的で大規模な実験データとして，任意のサイズの入力を自動生成可能な決定期学習のためのデータと，OpenImageと呼ばれる大規模な画像分類から得られた特徴データを利用した．これらのデータは100万から170万件の要素からなり，本研究で提案したシステムの実装はいずれの入力も扱うことが可能だった．さらにアクターモデルを利用しない既存のシステムと比べて，並列環境での改善は 1.4倍から3.1倍の速度向上を確認し，さらに2台の分散計算機環境でより多くのメモリを扱えることで100万件以上のデータの処理が可能であることを確認し，提案した拡張機能が大規模データに対しても実現可能であり，スループットの向上と既存の手法と同等の精度が得られることを確認した．

(9) 今後の展望としては，今回扱ったアルゴリズムに対する拡張をプログラミング言語の機構として開発することが考えられる．また，今回は直接扱わなかったが，実行履歴の分析にも具体的な応用が期待できる．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Sakurai Kouhei、Shimizu Taiki	４．巻 1
２．論文標題 Actor-based incremental tree data processing for large-scale machine learning applications	５．発行年 2019年
３．雑誌名 AGERE 2019: Proceedings of the 9th ACM SIGPLAN International Workshop on Programming Based on Actors, Agents, and Decentralized Control	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1145/3358499.3361220	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Taiki Shimizu, Kohei Sakurai
２．発表標題 Comprehensive Data Tree by Actor Messaging for Incremental Hierarchical Clustering
３．学会等名 IEEE COMPSAC2018（国際学会）
４．発表年 2018年

１．発表者名 櫻井孝平
２．発表標題 アクターモデルを適用した木構造データ処理のための状態共有を利用した負荷分散
３．学会等名 電子情報通信学会 知能ソフトウェア工学研究会
４．発表年 2017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
連携研究者	山根　智 (Yamane Satoshi) (70263506)	金沢大学・電子情報学系・教授 (13301)	