様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １２ 日現在

機関番号：８２６２６
研究種目：若手研究(B)
研究期間：2012～2014
課題番号：２４７００１１１

研究課題名（和文）ストリーム処理とデータ分析処理を統合した戦略的データ活用基盤の開発

研究課題名（英文）Development of Data Management Framework Integrating Stream Processing and Analytical Data Processing

研究代表者

油井　誠（Yui, Makoto）

独立行政法人産業技術総合研究所・情報技術研究部門・主任研究員

研究者番号：１０５８６７１２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、データベースとHadoopをハイブリッドに利用したスケーラブルな機械学習手法を開発した。バッチ学習をHadoop上で行い、逐次的な学習処理を関係データベースの一種であるPostgreSQL上で行う。KDD Cup 2012, Track 2の商用広告データセットを用いた回帰分析タスクで提案手法の有効性の評価を行い、State-of-the-artな機械学習フレームワーク（Vowpal Wabbit、Bismarck）等の比較を行い、Vowpal Wabbitに対して5倍、Bismarckに対して5倍から7.65倍の学習速度が得られるという結果を得た。

研究成果の概要（英文）：We proposed a database-Hadoop hybrid approach to scalable machine learning where batch-learning is performed on the Hadoop platform, while incremental-learning is performed on PostgreSQL.
We conducted a series of exterimental evaluation using a commercial advertisement dataset provided in the KDD Cup 2012, Track 2. The experimental results show that our scheme has a superior training speed compared with state-of-the-art scalable machine learning frameworks, 5 and 7.65 times faster than Vowpal Wabbit and Bismarck, respectively, for a regression task.

研究分野：データベース学

キーワード：機械学習　ビッグデータ　データベース　関係データベース　オンライン学習　確率的勾配降下法

１．研究開始当初の背景

Web データ、画像情報、センサデータ等の増加により大規模データを利活用しよう、データから価値を見出していこうという社会的な動向がある。今後データ管理基盤にはデータの価値化までが求められるおり、ビックデータの価値化では機械学習が鍵となる。

ビッグデータからの機械学習を利用したデータ駆動の意思決定において、事前に蓄積された訓練データに加えて Web アクセスのクリック軌跡データやリアルタイム事象を考慮し、より的確な意思決定を行うことが求められてる。こうした意思決定のプロセスには過去のデータを訓練事例として入力とする機械学習が頻繁に用いられる。

しかし、一般的に機械学習では学習が収束するまで学習を進める必要があるため、データ蓄積した上でのバッチ的な計算が不可欠である。時事刻々と生成されるストリームデータからのオンライン学習と蓄積された大量のデータからのバッチ的な学習を両立する必要がある。

２．研究の目的

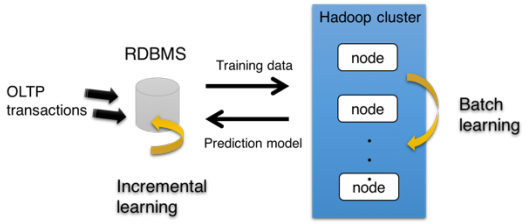
本研究の目的は、ストリーム型データ処理とバッチ型のデータ分析処理を統合することにより、速報性と正確性を両立したデータ駆動の戦略策定を支援するデータ活用基盤を開発することである。

３．研究の方法

(1) ハイブリッド学習

本研究では、データから逐次的に学習を進めるオンライン機械学習とバッチ的な機械学習を統合することで、蓄積されたデータからの高精度の学習モデルを構築することと、時事刻々と生成されるデータからの適応的な学習処理を両立させた。

具体的には、下記の図に示すとおりバッチ学習を MapReduce/Hadoop を用いて行い、逐次的（インクリメンタル）な学習を関係データベース上で行うハイブリッドなアプローチを採用した。



本提案手法では、Hadoop 上でのバッチ処理により全数データで安定した解の推定を行った上で、データベースの更新にも対応して新しいデータにも適応的に予測モデルをフィットさせていく。

下記の表に従来型のバッチ学習、オンライン学習とハイブリッド学習の予測精度、スループット、レイテンシを比較する。

	バッチ学習	オンライン学習
予測精度	○	△
スループット	数百万件/秒	数万件/秒
レイテンシ	数時間	数秒

	ハイブリッド学習
予測精度	◎
スループット	数百万件/秒
レイテンシ	数秒

ハイブリッド学習の利点は、高い予測精度を低いレイテンシ（遅延）で得ることができる点にある。このため、生成速度の速いデータソースからの学習に対応できる。

(2) 申請段階からの研究方法の変更点

交付申請書段階では、ストリーム処理に CEP 処理系を利用することを想定していたが申請者が入手可能なクリックストリームデータを取り扱う上で技術的に必要事項ではなかったため、優先事項から外し、大規模機械学習手法の開発に注力した。
「速報性と正確性を両立したデータ駆動の戦略策定を支援するデータ活用基盤」の実現手法と研究対象は、より機械学習寄りのものとなったが、ストリーム処理と蓄積型のバッチ処理のハイブリッドな手法という点で当初の研究目的に沿ったものとなっている。

４．研究成果

(1) 主な研究成果

スループット重視のバッチ学習は、Hadoop 上に構築した 32 台の並列学習器により約 2,300,000 tuples/sec のトレーニング速度を実現した。そして、レイテンシ重視のインクリメンタル学習は、PostgreSQL 上に実装し、70,000 tuples/sec のトランザクショナルな更新に対してインクリメンタルな学習モデルのメンテナンスを約 5 sec のレイテンシで行うことができることを確認した。

提案手法の評価として、KDD Cup 2012 の広告クリック率推定タスクを用いて State-of-the-art な機械学習フレームワーク(Vowpal

Wabbit、Bismarck)等の比較を行い、Vowpal Wabbit に対して 5 倍、Bismarck に対して 5～7.65 倍の学習速度が得られるという結果を得た。

単一計算機で動作するシングルプロセスの一般的な機械学習ライブラリでは 4 時間以上かかる処理を 32 台の計算機を利用した並列処理により 2 分以内に学習を完遂することが可能となった。

(2) 主な研究発表

ハイブリッドな機械学習手法を論文としてまとめ、ビッグデータ分野の主要な会の一つである IEEE 2nd International Congress on Big Data で発表を行った（学会発表⑤)。

開発成果をオープンソースソフトウェアの Hivemall として公開した。公開したソフトウェアに関する発表は、機械学習分野の最難関会議である NIPS のワークショップ（NIPS 2013 Workshop on Machine Learning Open Source Software）（学会発表④）や採択率２割をきる産業界からの注目度の高いエンジニアリングカンファレンスの Hadoop Summit 2014（学会発表③）に採択された。

また、第 26 回コンピュータシステム・シンポジウム（ComSys2014）における招待講演および、第 20 回先端的データベースと Web 技術動向講演会 （ACM SIGMOD 日本支部第 57 回支部大会）における依頼講演を行った。

最終的な成果をまとめた論文は、査読付き国内論文誌の情報処理学会論文誌：データベースに発表した。

(3) 研究アウトリーチ活動

研究成果の産業移転、アウトリーチ活動の一環として、オンライン広告関連会社の株式会社ロックオンと資金提供型の共同研究を進めた。テラバイト以上のデータ（月 50-70GB で増加中）の機械学習処理を実現し、企業側からのプレスリリースに至った。

http://www.lockon.co.jp/release/3074/

(4) 今後の展望

本課題の成果物として作成した機械学習ソフトウェアの Hivemall は既に国内外の複数の企業で利用されている。
今後も継続的に機能強化を行うことで、大規模機械学習を行う上でのデファクトスタンダードとしての地位を確立することを目指す。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① 油井誠, 小島功、Apache Hiveを用いたスケーラブルな機械学習機構の構築、情報処理学会論文誌：データベース, Vol. 8, No. 1, pp. 73-87, 情報処理学会, 2015年3月. 査読有
http://id.nii.ac.jp/1001/00079662/

〔学会発表〕（計６件）

① 油井誠. "Hivemall: Apache Hiveを用いたスケーラブルな機械学習ライブラリ", 第26回コンピュータシステム・シンポジウム（ComSys2014）,2014年11月19日. 芝浦工業大学豊洲キャンパス（東京都）

② 油井誠. "Hivemall: Apache Hiveを用いたスケーラブルな機械学習基盤", 第20回先端的データベースとWeb技術動向講演会 （ACM SIGMOD 日本支部第57回支部大会）, 2014年10月4日. リコーITソリューションズ株式会社本社事業所（東京都）

③ Makoto Yui. "Hivemall: Scalable Machine Learning Library for Apache Hive", 2014 Hadoop Summit, June 3-5, 2014. San Jose, CA, USA.

④ Makoto Yui and Isao Kojima. "Hivemall: Hive scalable machine learning library", NIPS 2013 Workshop on Machine Learning Open Source Software: Towards Open Workflows, Dec 10, 2013. Lake Tahoe, Nevada, USA.

⑤ Makoto Yui and Isao Kojima. "A Database-Hadoop Hybrid Approach to Scalable Machine Learning", IEEE 2nd International Congress on Big Data, June 27 - July 2, 2013, Santa Clara, CA, USA.

⑥ Steven Lynden, Isao Kojima, Akiyoshi Matono, Akihito Nakamura and Makoto Yui."A Hybrid Approach to Linked Data Query Processing with Time Constraints", The 6th Workshop on Linked Data on the Web (LDOW), May 14, 2013, Rio de Janeiro, Brazil.

〔その他〕
ホームページ等
https://github.com/myui/hivemall

６．研究組織

(1)研究代表者

油井　誠（MAKOTO YUI）
独立行政法人産業技術総合研究所・情報技術研究部門・主任研究員
研究者番号：10586712