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

平成　３０　年　　６　月　　７　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2014～2017
課題番号：　２６７３０１１４

研究課題名（和文）ビッグデータを用いた機械学習に適した最適化アルゴリズムとアーキテクチャの構成

研究課題名（英文）Architectures and optimization algorithms for machine learning from big data

研究代表者
松島　慎（Matsushima, Shin）

東京大学・大学院情報理工学系研究科・常勤講師

研究者番号：９０７２１８３７

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究では第一にSVMやロジスティック回帰などを包含する正則化付き経験リスク最初化問題について、複数のプロセスが非同期的に動作することで最適化を行うことができるスキームを提案、効率的な分散学習が行えることを理論・実験の両面から示した。
第二に、従来では数TBのデータを用いなければ学習できないスパース学習について、扱うデータ量を抑えながらスパース学習が可能であるスキームを提案した。提案手法はテキストデータやDNA配列データなどでは部分文字列の特徴量を用いた学習に関して、接尾辞配列などの効率的なデータ構造を用いる事によって、部分文字列に対応する特徴を効率よく抽出する事が可能であることを示した。

研究成果の概要（英文）：In this research, firstly, we proposed an optimization scheme for regularized empirical risk minimization that includes SVM and logistic regression. we have shown that this scheme that performs optimization by operating multiple processes asynchronously allows efficient distributed optimization from both theoretical aspects nd experimental aspect.
Secondly, focusing on sparse learning that originally requires several tera-bytes of data, we proposed an optimization scheme that works efficiently by suppressing the amount of data. We have shown that the proposed method can extract features efficiently by using efficient data structure such as suffix array in cases in which substrings are used as features of datasets such as text and DNA.

研究分野：　機械学習

キーワード：　機械学習　　凸最適化　　スパース学習　　大規模学習　　SVM

１．研究開始当初の背景

近年ビッグデータをいかに分析するかが情報学の一問題となっており、機械学習手法によってビッグデータを分析する事は非常に現実的かつ重要な課題である。時間的にも空間的にも高効率なパラメータ学習の方法を開発する事で、扱うことができるデータ量が増え、困難なタスクが学習可能となり、結果として今まで全く機械学習が有効でなかった分野での応用が爆発的に広がることも期待される。計算機は単に計算能力という側面だけでなく多角的な発展を遂げており、現在計算機は特有のメモリ階層構造を有し、マルチコアプロセッサや、Solid State Drive (SSD) などの高機能なメモリデバイスも利用可能になってきている。そのため、パラメータの学習の効率を高めるためには、用いる凸最適化アルゴリズムの理論的アプローチだけではなく、ハードウェアの構造も考慮されたスキーム全体としての時間効率に焦点を置くことが重要である。
特に、ほとんどの機械学習手法ではデータがメモリ容量よりも大きい場合に高速な学習が困難となるという問題を抱えている。メモリ容量よりも大きいデータを処理するためにはハードディスクや、ネットワークを介した各マシンへの不効率なアクセスを避けるという新たな視点が必須であり、申請者はデータサイズが RAM 容量を超える場合の効率的な SVM 学習スキームである Dual Cached Loops の提案・開発を行ってきた。

２．研究の目的

本研究について研究申請者ははビッグデータの性質と目的に適応した凸最適化アルゴリズムとアーキテクチャを構築することを目的とし、以下のように２つの研究目的を掲げた。

(1). SSD を用いた単一マシンにおける Dual Cached Loops のさらなる大規模化
研究申請者がこれまでの研究で提案した Dual Cached Loops を拡張し HDD やネットワーク上のデータを、SSD と RAM の２層の転送速度の異なるデバイスを用いて、それぞれの転送速度の違いを利用した SVM 学習のための最適化アルゴリズムを設計する。これにより、単一のマシンで数十 TB 単位の大規模データを用いた学習が可能であ

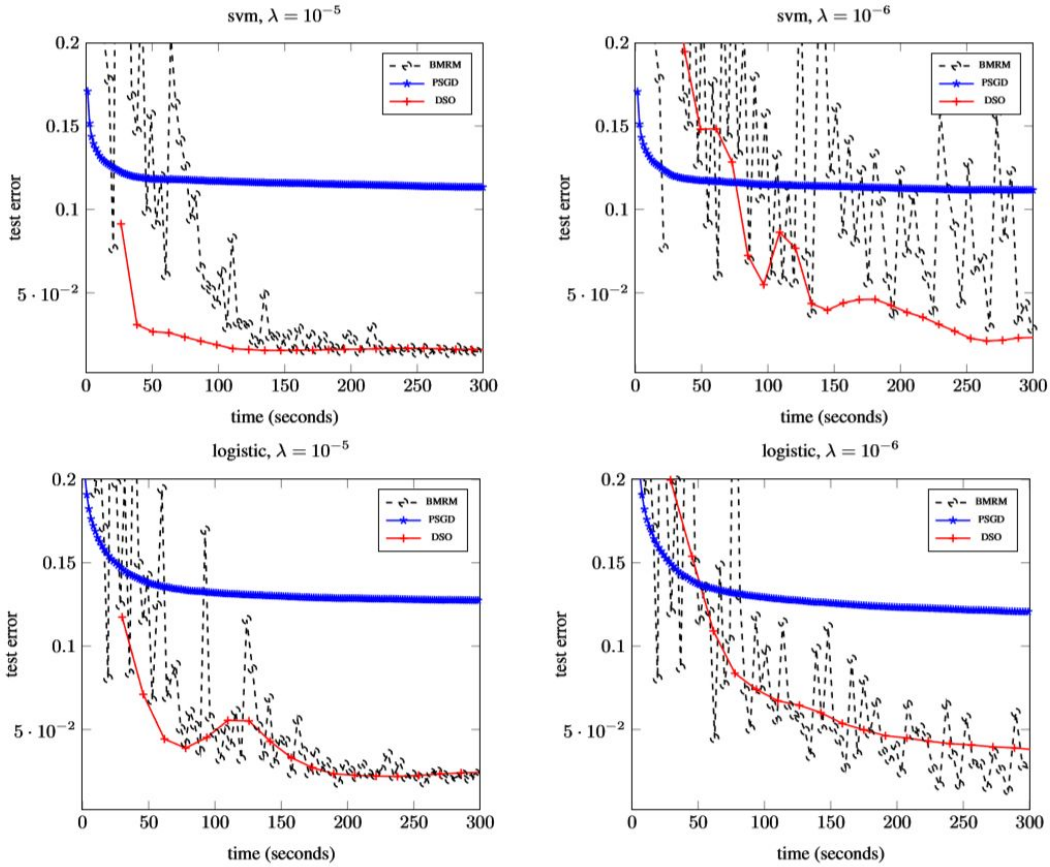

図 1　DCL のさらなる大規模化における提案スキーム（図中 DSO）を代表的な既存手法である分散確率的勾配法（図中 PSGD）やバンドル法（図中 BMRM）と比較した結果を示す。４つのパネルはそれぞれの最適化問題において各手法の経過時間、テストデータにおける識別誤差を表す。上部は SVM を用いた比較、下部はロジスティック回帰を用いた比較である。

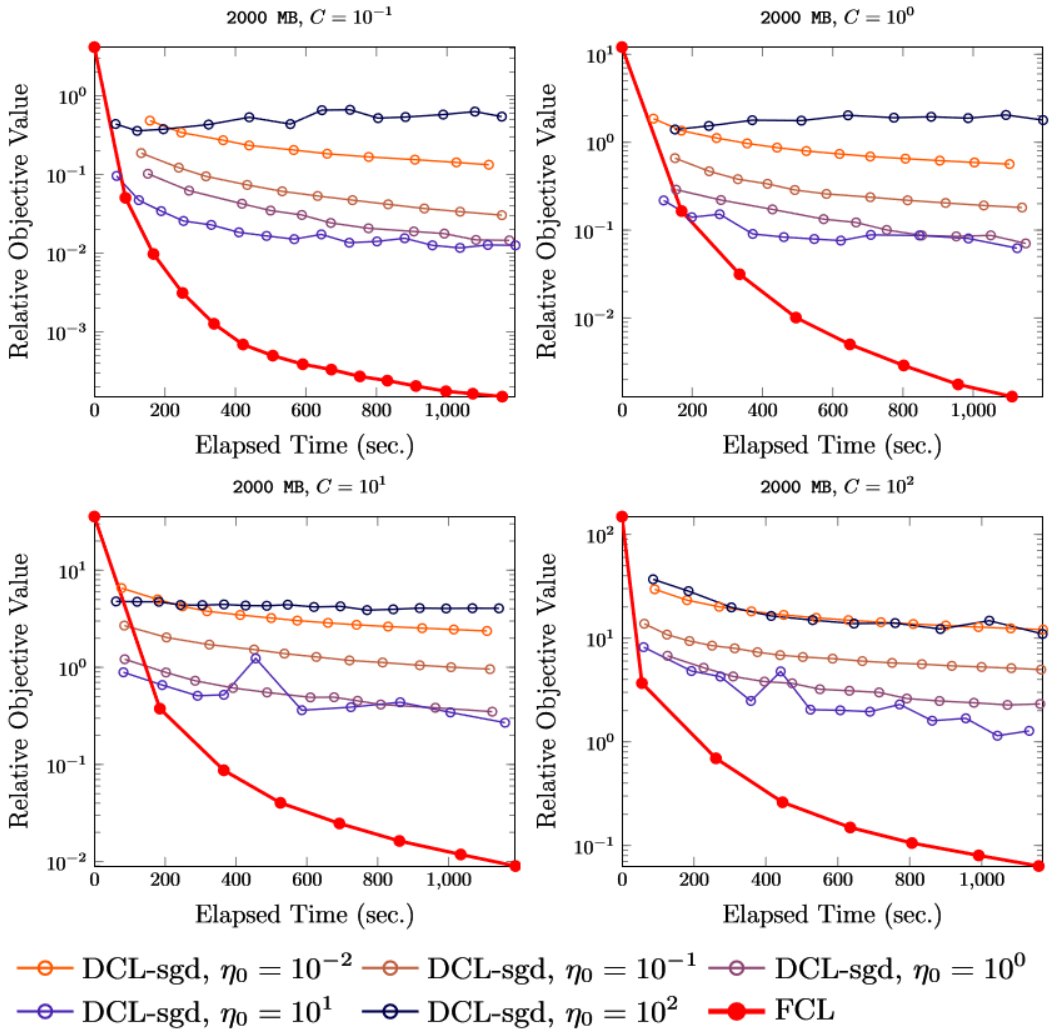

図 2　スパース学習における提案手法（図中 FCL）を既存手法（図中 DCL-sgd）と比較した結果を示す。4つのパネルはそれぞれの最適化問題において各手法の経過時間、テストデータにおける最適化誤差を表す。

ることを明らかにする。

(2).スパース学習を用いた Dual Cached Loops の拡張

スパース学習において従来では非常に大きなデータ行列を処理しなければならない場合にも、Dual Cached Loops を用いてデータ行列を列ごとに生成するスレッドと最適化を行うスレッドを非同期的に動作させ、大規模データを直接扱うことなく同等のデータに対する最適化が可能なスキームを構築する。これにより、単一のマシンで数 TB 単位の大規模データを用いたスパース学習が可能であることを明らかにする。

3．研究の方法

(1). SSD を用いた単一マシンにおける Dual Cached Loops のさらなる大規模化

SSD の特性を生かすことにより数十 TB のデータに基づく SVM の学習が可能であるスキームを考案、実装する。さらに、それを実際に利用した知識発見の応用を行い、開発されたスキームの有効性を示し、実データに応用し実際に有用な知識を取り出す事を目的とした実践的なデータマイニングとしての可能性を探る。

(2).スパース学習を用いた Dual Cached Loops の拡張

従来では数 TB のデータを用いなければ学習できないスパース学習について、扱うデータ量を抑えながらスパース学習が可能であるスキームを考案、実装する。さらに、それを実際に利用した知識発見の応用を行い、開発されたスキームの有効性を示す。特に研究の後半段階では、本来は非線形識別でしか達成できないほどの高次元写像を利用可能なアルゴリズムを検討する。

4．研究成果

(1). SSD を用いた単一マシンにおける Dual Cached Loops のさらなる大規模化

SVM だけでなく、ロジスティック回帰などを包含する正則化付き経験リスク最初化問題について、複数のプロセスが同時に動作することで最適化を行うことができるスキームを考案、実装した。本スキームは単一マシンにおいても複数マシンにおいても同様に動作することができる汎用的なものである。図１は提案スキームを代表的な既存手法である確率的勾配法やバンドル法と比較した結果を示す。図中上部は SVM を用いた比較、下部はロジスティック回帰を用いた比較であり、既存手法より効率的にパラメータ学習を行うことができた。本成果に基づいた論文は ECML-PKDD にて採択され発表を行った。

(2).スパース学習を用いた Dual Cached Loops の拡張

従来では数 TB のデータを用いなければ学習できないスパース学習について、扱うデータ量を抑えながらスパース学習が可能であるスキームを考案、実装した。特にテキストデータや DNA 配列データなどでは部分文字列の特徴量を用いた学習により、接尾辞配列などの効率的なデータ構造を用いる事によって、部分文字列に対応する特徴を効率よく抽出する事が可能であることを示した。図２は提案スキームと既存の Dual Cached Loops と比較した結果を示す。提案スキームは既存手法よりも効率的に最適解に近づくことが示された。本成果に基づいた論文は ECML-PKDD にて採択され、発表を行った。

５．主な発表論文等

〔雑誌論文〕（全て査読あり）（計２件）

(1) S. Matsushima, H. Yun, X. Zhang, S.V.N. Vishwanathan. “Distributed Stochastic Optimization of Regularized Risk via Saddle-Point Problem.” *In Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases* (ECML PKDD), *Lecture Notes in Computer Science* (LNCS) 10534, pp 460-476, (2017)

(2) S. Matsushima. “Asynchronous Feature Extraction for Large-scale Linear Predictors." *In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (ECML PKDD), *Lecture Notes in Computer Science* (LNCS) 9851, pp 604-618, (2016)

〔学会発表〕（計３件）

(1) S. Matsushima, H. Yun, S.V.N. Vishwanathan. 「正則化付き経験リスク最小化における分散最適化法」日本応用数理学会 2014 年度年会, 東京, 2014 年 9 月

(2) 松島慎.「大規模な線形予測器のための非同期特徴抽出スキーム」統計的モデリングと計算アルゴリズムの数理と展開, 愛知, 2017 年 2 月

(3) 松島慎.「大規模な L1 正則化問題のための座標降下法を用いたスキーム」第 31 回人工知能学会全国大会, 愛知, 2017 年 5 月

〔その他〕

ホームページ等

https://ml.c.u-tokyo.ac.jp/research

６．研究組織

(1) 研究代表者

松島　慎（MATSUSHIMA, Shin）

東京大学・大学院情報理工系研究科・常勤講師

研究者番号：90721837