Incremental and Parallel Learning Algorithms for Data Stream Knowledge Discovery*

Lei Zhu

Abstract

Incremental and parallel are two capabilities for machine learning algorithms to accommodate data from real world applications. Incremental learning addresses streaming data by constructing a learning model that is updated continuously in response to newly arrived samples. To solve the computational problems posed by large data sets, parallel learning distributes the computational efforts among multiple nodes within a cloud or cluster to speed up the calculation. With the rise of BigData, data become simultaneously large scale and streaming, which is the motivation to address incremental and parallel incremental (PI) learning in this work.

This research first considers the incremental learning alone, in the graph max-flow/min-cut problem. An augmenting path based incremental max-flow algorithm is proposed. The proposed algorithm handles graph changes in a chunking manner, updating residual graph via augmentation and de-augmentation in response to edge capacity increase, decrease, edge/node adding and removal. The theoretical guarantee of our algorithm is that incremental max-flow is always equal to batch retraining. Experiments show the deterministic computational cost save (i.e., gain) of our algorithm with respect to batch retraining in handling graph edge adding.

The proposed incremental max-flow is then applied to upgrade an existing batch semi-supervised learning algorithm known as graph minicuts to be incremental. In batch graph minicuts, a graph is learned from input labeled and

*Doctoral Dissertation, Graduate School of Information Science,
Nara Institute of Science and Technology, NAIST-IS-DD123456, February 1, 2018.

unlabeled data, and then a min-cut is conducted on that graph to make the classification decision. In the proposed modification, the graph is updated dynamically for accommodating online data adding and retiring. Then the proposed incremental max-flow algorithm is adopted to learn min-cut from the resulting non-stationary graph. Empirical evaluation on real world data reveals that the proposed algorithm outperforms state-of-the-art stream classification algorithms.

In the incremental max-flow, the training speed is not satisfactory when the data set is huge. A straightforward solution is to combine parallel data processing with incremental learning. Previously, parallel and incremental learning are often treated as two separate problems and solved one after another. Alternatively in this work, these two learning problems are solved in one process (i.e., PI integration).

To simplify the learning, this research considers a base model in which incremental learning can be implemented by merging knowledge from incoming data and parallel learning can be performed by merging knowledge from simultaneous learners (i.e., in knowledge mergeable condition). As a result, this work develops a parallel incremental wESVM (weighted Extreme Support Vector Machine) algorithm, in which the parallel incremental learning of the base model is completed within a single process of knowledge merging. Specifically, the wESVM is reformulated such that knowledge from subsets of training data can be merged via simple matrix addition. As such, the proposed algorithm is able to conduct parallel incremental learning by merging knowledge from data slices arriving at each incremental stage. Both theoretical and experimental studies show the equivalence of the proposed algorithm to batch wESVM in terms of learning effectiveness. In particular, the algorithm demonstrates desired scalability and clear speed advantages to batch retraining.

In the field of data stream knowledge discovery, this work investigates incremental machine learning and invents a wESVM-based parallel learning and incremental learning integrated system. The limitation of this work is that PI integration applies only to models that satisfy the knowledge mergeable condition. Future work should investigate how to release this constraint and expand PI integration to other models such as SVM and neural network.

氏　名	Lei Zhu

（論文審査結果の要旨）

ビッグデータ時代を迎え，日々増加する巨大なデータを効率的に処理する方法が求められている．そのための機械学習の手法に，追加学習および並列学習がある．本研究は従来別々に研究されていたこれらの要素を統合し，効率の良い学習方法を提案したものである．

本論文ではまずはじめに，max-flow/min-cut 問題のための追加学習法を提案している．通常，max-flow 問題ではネットワークの変化について考慮しないが，現代の情報ネットワークは時間変化することからそれに対応する必要がある．そこで本論文では，バッチ学習と同じ解を与えることが理論的に保証された追加学習法を提案した．さらに，この手法は准教師あり学習と組み合わせることで min-cut 問題にも適用できることを示した．
次に追加学習法を並列学習と組み合わせる方法を提案している．ここでは学習機械として wESVM (weighted Extreme Support Vector Machine) を利用することで，新規データをマージする追加学習と他学習機の結果をマージする並列学習を効率的に組み合わせることに成功した．さらにこの学習法がバッチ学習と同じ解を与えることも理論的に証明している．

以上をまとめると，本論文はビッグデータ時代の情報処理に適した追加学習及び並列学習を統合した学習アルゴリズムを提案しており，ビッグデータ解析の発展に資すると考えられる．よって，博士（工学）の学位に値するものと認められる．