並列分散環境における Random Forests の学習法に関する研究

若山 涼至　　指導教授：藤吉 弘亘

1. はじめに

本研究では，Random Forests(RF)[1] を利用した画像処理のための分類・回帰の研究について取り組んだ．本稿では，並列分散処理における RF の学習法について述べる．

データの増加に伴い，機械学習の学習時間は増加している．この問題に対して，複数の計算機で並列に処理することで処理時間を短縮する取り組みが行われている．複数の計算機で処理を行う場合，効率性から計算機間でデータの重複がないように分割されることが一般的である．機械学習において，データを分割した際にバイアスが発生すると過学習により分類性能が低下する問題がある．そこで，本研究では共変量シフトを導入した Random Forests[2] を用いてデータ分割時に発生するバイアスの影響を低減する学習法を提案する．

2. MapReduce[3]

MapReduce は並列分散処理を効率的に行うために提案されたモデルである．このモデルは，全ての計算機を統括する計算機 (マスタノード) と実際に処理を行う計算機 (ワーカノード) から構成される．

Map 処理

Map 処理では得られたデータをマスタノードへ与え，マスタノードはワーカノードへデータを割り振る．ワーカノードは得られたデータに処理を行い，処理結果をマスタノードへ返還する．

Reduce 処理

Map 処理後，マスタノードは返還された処理結果を統合し，解決すべき問題の答えを出力する．

この処理により膨大な量のデータに対し比較的少ない時間で処理を行うことが可能となる．

3. 並列分散学習における Random Forests の学習

本研究では，図 1 に示す 2 つの MapReduce のモデルにおける学習法を提案する．本稿では，図 1(a) のモデルをマスタノード非通過型，図 1(b) のモデルをマスタノード通過型モデルと定義する．提案手法では，共有データを導入し，RF の枠組みで転移学習を行う Transfer Forests を利用する．以下で Transfer Forests について説明し，次に MapReduce による RF の学習について説明する．

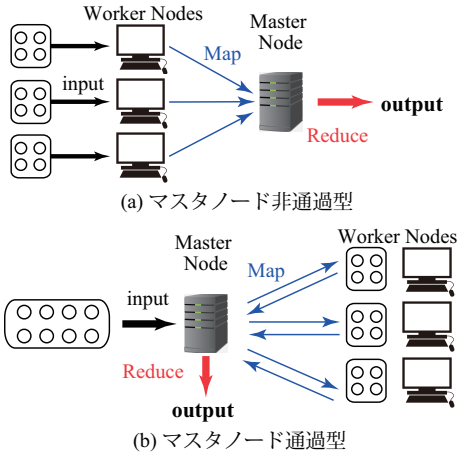

図 1：MapReduce のモデル

3.1 Transfer Forests

Transfer Forests は RF のフレームワークで転移学習を行う手法である．図 2 に Transfer Forests の学習の概要を示す．転移学習では，既に得られているサンプルや分類器を事前ドメインとし，それらを利用して目標サンプルの学習を行う．目標サンプルを学習する際に共変量シフトを利用し，学習への有効性により事前サンプルへ重み付けする．この重み付けによって学習に有効でないサンプルに対して，目標サンプルへの学習に影響を与えにくくする．共変量 λ は以下により求められる．

$$\lambda_j = \frac{1 + e^{P_s}}{1 + e^{P_t}} \tag{1}$$

ここで，P_s，P_t はそれぞれ事前ドメインと目標ドメインの強分類器の事後確率を表す．

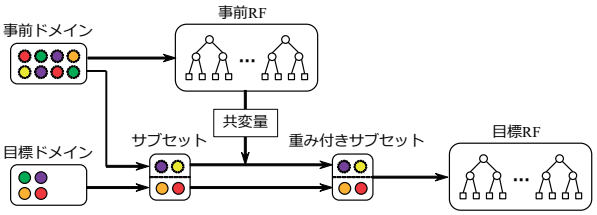

図 2：Transfer Forests

3.2 Map 処理

本研究では，マスタノード通過型とマスタノード非通過型の 2 つのモデルで学習を行う．これら 2 つのモデルでは Map 処理が異なるため，以下でそれぞれ説明する．

マスタノード非通過型

マスタノード非通過型では，データはワーカノードへ直接入力，配置が行われる．そのため，マスタノードによるデータ調整が不可能であるために，データの分布の偏りが発生する場合がある．RF を学習する場合，学習データの分布に偏りが発生すると過学習により分類性能が低下する問題がある．

そこでマスタノード非通過型では，ワーカノードが持つデータとは別に，全てのワーカノードで共有可能なデータを導入することによりデータの分布の偏りを抑制する．また，共有データからワーカノードが持つデータの学習に有効なデータを利用するために，転移学習により学習を行う．この時，共有データを転移学習の事前サンプル，ワーカノードが持つデータを目標サンプルとして学習する．そのため，共有データからあらかじめ事前 RF が構築されているものとする．ワーカノードは Transfer Forests により目標 RF を構築し，マスタノードへ返還する．

マスタノード通過型

マスタノード非通過型では，データ配置時に分布の調整を行うことができないために，分布の偏りが発生し過学習を引き起こす問題がある．しかし，マスタノード通過型においては，マスタノードが配置先のワーカノードを指定可能であることから，サンプルの分布に偏りが発生することはない．そのため，共有データを導入する必要はなくデータ学習の処理においては通常の RF により効率的に学習を行うことができる．

しかし，データの分布が偏らない場合においても，ワーカ数が増えることにより 1 つのワーカノードに対して学習に必要最低限なデータ量を確保することが難しくなる．この問題に対しても，共有データの導入が有効である．共有データを導入することにより全てのワーカノードに一定数のデータが与えられるため，学習に必要最低限のデータ量を確保することが可能である．マスタノード通過型も非通過型と同様に，構築した RF をマスタノードへ返還する．

3.3 Reduce 処理

マスタノードは並列に構築された目標 RF を集約し，1 つの決定木群を生成する．並列分散処理により過剰に決定木を構築した場合には分類時に余分なコストを生み出す．そのため，決定木を削除することで分類時の計算コストを軽減する．

決定木の削除には，乱数を用いて削除する方法や，スコア算出用サンプルによりスコアを算出し決定木を削除する方法を用いる．乱数による決定木の削除では，RF のランダム性を損なうことなく決定木を削除することができ，また決定木を少ない計算コストで削除することが可能である．ワーカノード毎に保持するサンプル数が異なる場合や，サンプルにノイズが多く含まれるような場合には決定木の性能に差が出やすいためにスコアによる決定木の削除を行うことで効率よく精度の高い決定木を残すことができる．本研究では，スコアは以下により算出する．

$$Score(t) \quad = \quad -\sum_{i=1}^{I} \max_{c \neq c_i} h(\mathbf{v}_i, c, t) \qquad (2)$$

ここで，I はサンプル数，$\mathbf{v}_i, c_i$ は，それぞれサンプルの特徴ベクトル，ラベルを表し，h は決定木 t が出力するクラス c の事後確率を表す．スコア関数は問題設定により変更することで様々な問題に適用することが可能である．Reduce の処理はマスタノード通過型，非通過型共に同じ処理によって行われるが，スコア算出に利用するサンプルは MapReduce のモデルに合わせて変更する．

4. 評価実験

提案手法の有効性を評価するために評価実験を行う．本実験では，Map 処理の評価のためにデータの分割時にバイアスが発生した場合の分類誤差を評価し，Reduce 処理の評価のために，乱数による決定木の削除とスコアによる決定木の削除による分類性能の変化の調査を行う．

4.1 データベース

本実験では，UCI Machine Learning Repository letter recognition を用いて評価を行う．letter recognition はサンプル数 20000，クラス数 26，特徴次元 16 のデータセットである．本実験では，6666 個を評価用サンプルとし残りを学習用サンプルに用いる．事前サンプルの数は 4000 個とする．決定木のパラメータは，本数 50，深さ 15，特徴次元選択回数 15 回，閾値選択回数 50 回とする．

4.2 実験 1：Map 処理の評価

本実験では，マスタノード非通過型 (MN 非通過型) とマスタノード通過型 (MN 通過型) における学習の評価を行う．マスタノード非通過型では，データ分布の調整が不可能であるために，データ分布に偏りがあるものとする．マスタノード通過型では，データ分布の調整が可能であるため，データ分布に偏りがないものとする．比較対象として共有データを用いない RF による学習 (RF) と，共有データを用いた転移学習 (Transfer Forests) を比較する．実験結果を表 1 に示す．実験結果から，マスタノード非通過型において，共有データを利用しない学習法では大きく分類性能が低下している．しかし，共有データを利用することで，データに分布の偏りが発生した場合においても精度の低下を抑制している．マスタノード通過型は，分布の偏りがないため通常の Random Forests で学習することにより共有データを必要としないが，マスタノード非通過型の場合には共有データを導入することが必要である．また，マスタノード通過型においても，ワーカ数が増えるにつれて 1 つのワーカノードに割り当てられるデータ量が少なくなり，決定木の分類性能が低下するため，ワーカ数が多い場合には共有データを利用することで決定木の分類性能の低下を抑制可能である．

表 1：分類誤差 [%]

	MN非通過型		MN通過型	
ワーカ数	RF	TF	RF	TF
1	6.19	5.14	6.19	5.14
2	9.82	5.71	7.78	5.83
5	26.4	8.28	10.6	7.29
10	67.72	9.53	18.3	8.59

4.3 実験 2：Reduce 処理の評価

本実験では，ノイズの割合を変化させて学習した決定木を削除したときの分類誤差を評価する．このときスコア算出には評価用サンプルの半分を利用し，評価用サンプルにはクラスラベルにノイズがないものとする．本実験では，ワーカ数を 10，木の削除本数を 20 とする．比較手法として，乱数による決定木の削除とスコアによる決定木の削除を比較する．結果を表 3 に示す．実験結果から，ノイズを含むような場合にはスコア算出により効率的に決定木を削除することが可能である．また，決定木を多く削除したときに，スコア算出により決定木を削除することで，乱数で削除した場合よりも比較的高い分類性能になる傾向が得られた．ノイズの割合が少ない場合や，決定木の削除数が少ない場合には，乱数で決定木を削除した場合においてもスコア算出による決定木の削除との差が小さいため，乱数により決定木を削除することで計算コストを小さくすることが可能である．

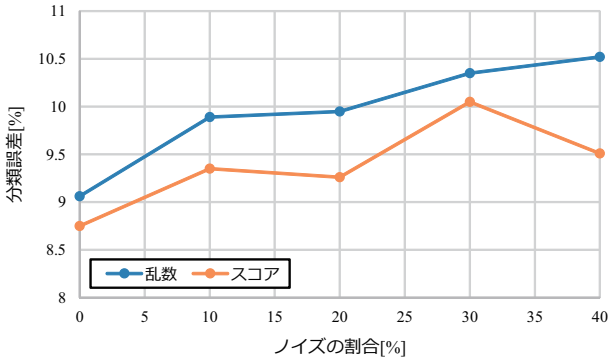

図 3：ノイズの割合変化時の分類誤差

5. おわりに

本研究では，Transfer Forests による並列分散環境の学習方法を提案した．Map 処理としてマスタノード非通過型，マスタノード通過型の 2 つの MapReduce のモデルにおける学習法を提案し，それぞれが持つ特徴について述べた．データの分布に偏りが発生する問題や，分割によりデータ量が不足する場合に対し共有データを利用することが有効であることを示した．また，Reduce 処理として決定木を削除する方法を提案し，効率的に決定木を削除することを可能としたことで，分類時の計算コストを削減した．共有データとスコア算出用データにより分類性能が大きく左右される．そのため今後は，共有データとスコア算出用データの選択法を確立することで，より高精度で効率的な並列分散学習を目指す．

参考文献

[1] L. Breiman, Random Forests, Machine Learning, vol.45, pp.5-32, 2001.

[2] 土屋成光，弓場竜，山内悠嗣，山下隆義，藤吉弘亘，共変量シフトに基づく Transfer Forest，信学技報，vol.114, pp.31-36, 2014.

[3] J. Dean, and S. Ghemawat, Mapreduce: simplified data processing on large clusters, Communications of the ACM, vol.51, no.1, pp.107-113, 2008.

研究業績

[1] 若山涼至, 藤吉弘亘,“複数パスを考慮した Regression Forests によるカメラのヨー角の推定”, PRMU 研究会, 2013.

[2] 倉貫芳紀, 若山涼至, 吉田睦, 藤吉弘亘,“移動物体の影響を低減した単眼モーションステレオによる距離推定”, 画像センシングシンポジウム, 2013.

（他 学会口頭発表 2 件　学会口頭発表予定 1 件）

受賞

[1] PRMU2013 ポスター賞

[2] 第 17 回 PRMU アルゴリズムコンテスト 最優秀賞