

# 知能システム学特論 最終レポート

Hdp 第2班 16344217 津上祐典

2016年8月19日

## 1 テーマ

Spark と Hadoop を用いた分散機械学習によるクラス分類 -スパムメールの検出と画像認識

## 2 概要

はじめに、Hadoop, Spark, 機械学習の原理について述べ、最後に実行結果と考察を示す.

### 2.1 Hadoop

Hadoop とはビッグデータを複数の PC を用いて分散並列処理を可能にするフレームワークである. 一台マスターサーバとその配下にある複数のスレーブサーバによって分散並列処理を行う. Hadoop は分散ファイルシステム (Hadoop Distributed File System), 並列分散処理フレームワーク (MapReduce Framework) より構成されている. 分散ファイルシステムとは複数のスレーブサーバを一つのストレージとして扱うファイルシステムである. 分散並列処理フレームワークでは与えられたデータから欲しいデータの抽出と分解する Map 処理, それらのデータを集計する Reduce 処理が行われる. MapReduce 処理を複数のスレーブサーバで行うことで分散処理を可能にし, ビッグデータを効率よく扱うことができる. Hadoop は分散並列処理システムであり, Hadoop のみでは機械学習が行えない. しかし, 機械学習するためのライブラリ等のツールがいくつか用意されている. Hdp 第2班では, Spark を用いた. 次節にて, Spark について説明する.

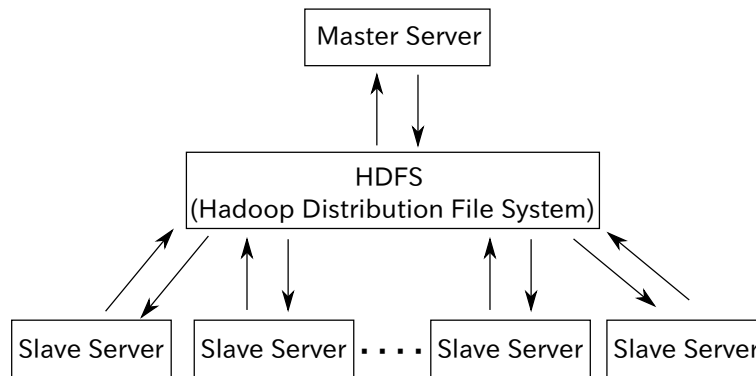


図 1. Hadoop の構成

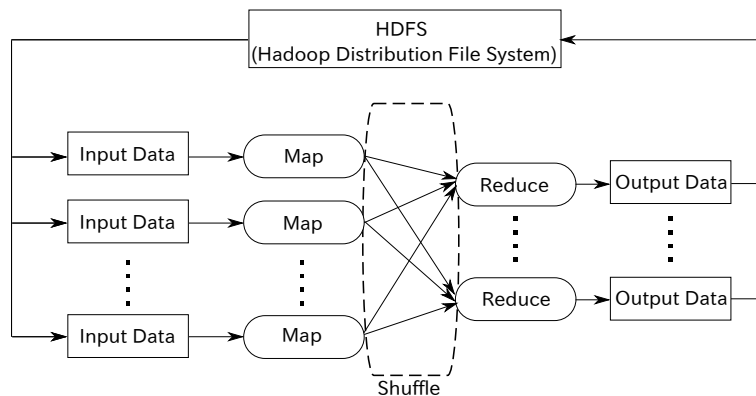
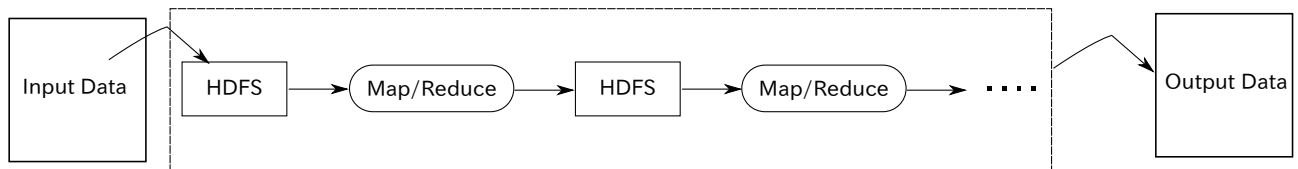


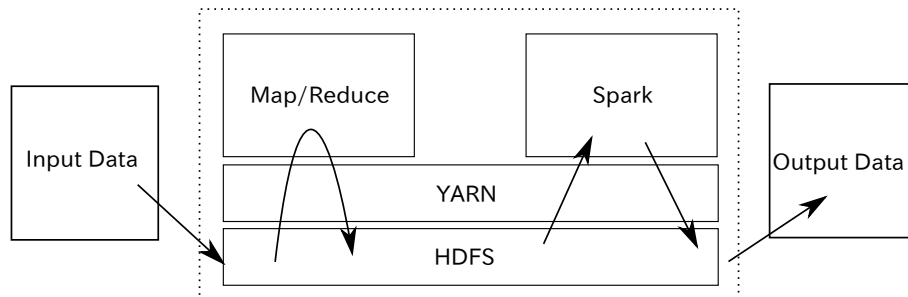
図 2. Hadoop の分散処理の流れ

## 2.2 Spark

Spark とは Hadoop 同様，分散並列処理を可能にするフレームワークである．Spark 自身は HDFS を持っておらず，Hadoop の HDFS を利用することが出来る．Hadoop は一つの処理が終わるたびに HDFS に書き込まなければならないが，Spark ではインメモリを用いることで一つの処理ごとに HDFS に書き込む必要が無く，処理を高速化している．図 3 に Hadoop の場合と Spark と Hadoop を組み合わせた場合の処理の流れを示す．図 3(b) では，まず，大きなデータを MapReduce で加工し，その後 Spark で分散処理している．Spark のデータ処理には RDD(Resilient Distributed Dataset) と呼ばれるデータ構造を用いている．RDD は大量のデータを要素として保持する分散コレクションである．RDD 内はパーティションというかたまりに分割されており，これが分散処理の単位である．RDD をパーティションごとに複数の PC で分散処理することで一台の PC では難しいビッグデータの分散処理が可能となる．また，機械学



(a) Hadoop のみ



(b) Hadoop と Spark の組み合わせ

図 3. Hadoop と Spark

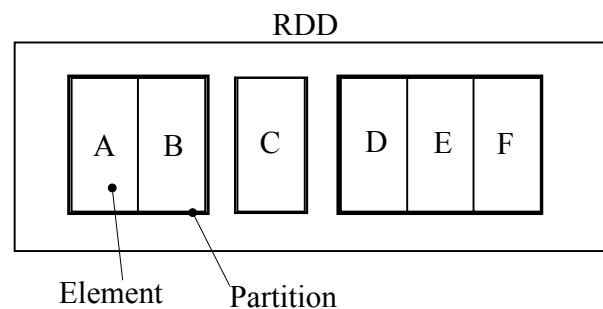


図 4. RDD の構造

習用ライブラリ MLlib が用意されており，機械学習の分散並列処理が可能である．使用可能な言語は Java, Python, R 言語である．

## 2.3 機械学習 (Machine Learning)

### 2.3.1 スパムメールの分類

一つ目の学習テーマとしてスパムメールの分類を行った．データセットとして Spambase DataSet[1] を用いた．このデータセットは 4601 通のメールがあり，うち 1813 通のスパムメールと 2788 通の非スパムメールから構成されている．また，57 次元のベクトルとして特徴量抽出済みである．1～48 番目の要素は特定の変数名の出現頻度，49～54 番目の要素は記号文字の

出現頻度，55～57 番目の要素は，大文字の連なりの長さの平均，最長，合計である．学習アルゴリズムとして，ロジスティック回帰を使用した．ロジスティック回帰とは，識別関数としてシグモイド関数を用いた回帰モデルである．シグモイド関数は以下の式で表される．また，図 5 にシグモイド関数のグラフを示す．

$$f_{\theta}(x) = \frac{1}{1 + e^{-\theta x}} \quad (1)$$

ただし， $\theta$  はパラメータである．目的関数である対数尤度関数  $\log L(\theta)$

$$\log L(\theta) = \sum_{i=1}^n (y^{(i)} \log f_{\theta}(x) + (1 - y^{(i)}) \log(1 - f_{\theta}(x))) \quad (2)$$

を最大化するようなパラメータ  $\theta$  を学習（更新）する．パラメータの更新式を求めるには，最急降下法や確率的勾配法，準ニュートン法などがあるが，Spark の MLlib で用意されている準ニュートン法を用いてパラメータを更新した．

### 2.3.2 画像の分類

二つ目の学習テーマとして画像の分類を行った．データセットとして CIFAR-10[2] を用いた．CIFAR-10 は 60000 枚の画像からなり，うち 50000 枚が訓練用画像であり，残り 10000 枚はテスト用画像である．オープンライブラリの scikit-image を HOG 特徴量を抽出し，前節と同じくロジスティック回帰で画像認識（他クラス分類）を行った．

## 2.4 実行条件

前述した 2 つのテーマの実行条件を表 1 に示す．

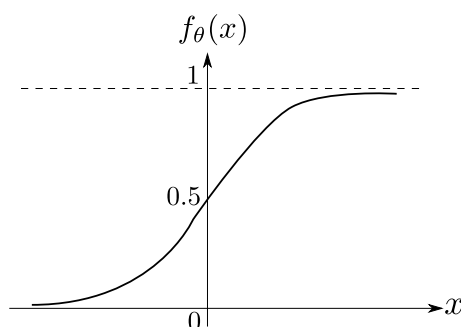


図 5. シグモイド関数

表 1. 実行条件

テーマ	スパムメールの検出	画像認識
データセット	Spambase Data Set	CIFAR-10
学習アルゴリズム	ロジスティック回帰	
評価法	ホールドアウト法	
環境	Master : 1 台, Slave : 2 台	
OS	Ubuntu 14.04 LTS	

表 2. スパムメールの検出の結果

非スパム再現率	スパム再現率	非スパム適合率	スパム適合率	AUC(PR)
91.92%	90.41%	93.48%	88.21%	0.9123

### 3 自分の分担範囲

機械学習・分散処理（Hadoop, Spark）の理論調査，毎週のレポート作成

### 4 感想

本プロジェクトを通して，ビッグデータの分散並列処理の原理を理解し，実装することができた．また，機械学習の理論を理解し，分散並列処理に適用できた．

### 5 評価

### 参考文献

- [1] "Spambase Data Set", <https://archive.ics.uci.edu/ml/datasets/Spambase>, 2016 年 8 月 9 日最終確認.
- [2] "The CIFAR-10 dataset", <https://www.cs.toronto.edu/~kriz/cifar.html>, 2016 年 8 月 9 日最終確認.

表 3. 画像認識の結果

	Airplane	automobile	Bird	Cat	Derr	Dog	Frog	Horse	Ship	Truck
Precision	55.53%	64.49%	42.98%	38.21%	42.69%	45.91%	51.63%	57.36%	57.91%	62.30%
Recall	55.70%	66.10%	35.80%	29.80%	42.90%	43.80%	63.40%	60.40%	62.20%	65.10%

表 4. 評価

学籍番号	氏名	評価
16344217	自分	機械学習の理論，プログラムを担当した．機械学習や分散処理の理論の方は理解できたが，プログラムの方ではあまり貢献できなかった．
16344201	井上 聖也	機械学習のプログラムを作ってくれた．理論やプログラムに関して色々教えてもらった．
16344216	田中 良道	Hadoop と Spark について調査してくれた．通信エラーの解決法など色々調査してくれた．
15344229	沈 歩偉	Hadoop,Spark の完全分散処理環境を構築してくれた．実行時間の比較実験などしっかりやってくれた．また，Hadoop,Spark のインストールに困っている時，手伝ってくれた．