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,Hadoop を用いた機械学習によるスパムメールの分類と画像の多クラス分類

2　概要

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

2.1　Hadoop

Hadoop 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．身近な例でクラウドやネットワーク HDD(NAS) などがある．並列分散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．

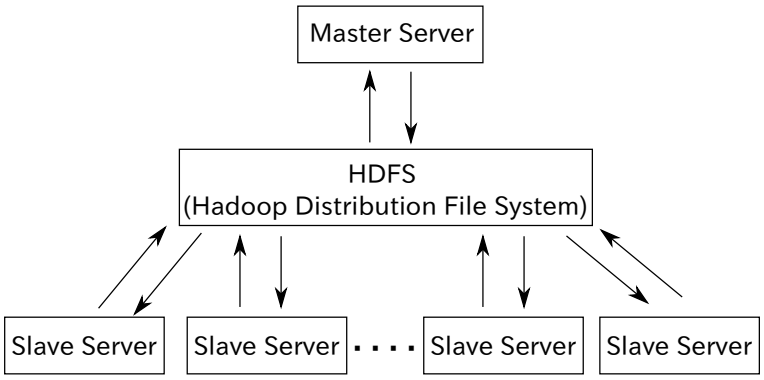

図 1. Hadoop の構成

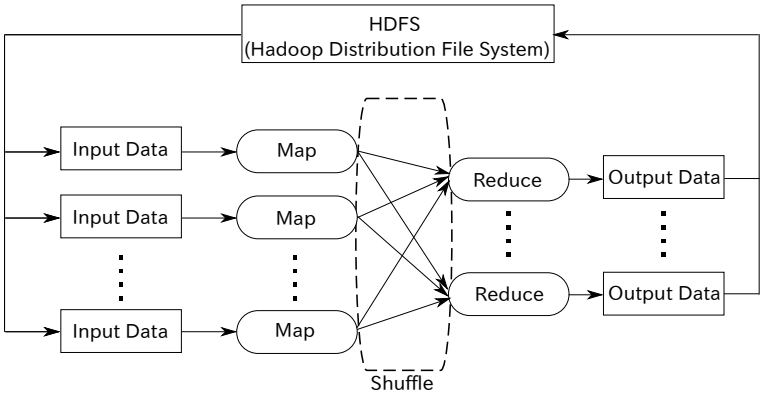

図 2. Hadoop の分散処理の流れ

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．Spark では機械学習用ライブラリ 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．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 ではインメモリに書き込むことで，一つの処理ごとに HDFS に書き込む必要が無く，繰り返し処理に強い特長がある．また，機械学習用のライブラリとして MLlib が用意されている．Spark は RDD(Resillient Distributed Datasets)

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 3 にシグモイド関数のグラフを示す．

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

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

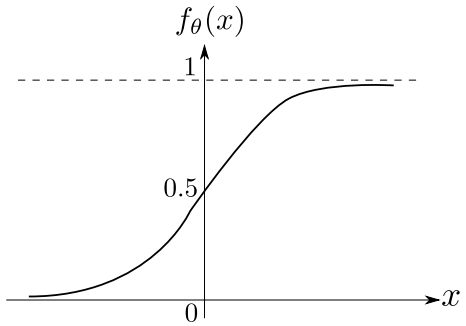

図 3. シグモイド関数

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

を最大化するするようなパラメータ θ 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．

2.4 実験

その結果を表 2 に示す．また，ロジスティック回帰モデルで設定パラメータを実験的にチューニングして求めた際の結果を表 3 に示す．

2.4.1 実行条件

表 1. 実行条件

テーマ	スパムメールの検出	画像の分類
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. 実験結果

	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)
SVM	76.13%	85.79%	88.90%	70.61%	0.8105
ロジスティック回帰	91.92%	90.41%	93.48%	88.21%	0.9123
ナイーブベイズ	83.85%	66.25%	78.79%	73.28%	0.7652

表 3. パラメータのチューニング前後の比較

	チューニング前	チューニング後
正則化関数	L2 ノルム	L1 ノルム
正則化係数	0.01	0.002
学習繰り返し回数	10	25
非スパム再現率	88.60%	91.92%
スパム再現率	88.81%	90.41%
非スパム適合率	92.21%	93.48%
スパム適合率	83.89%	88.21%
ACU(PR)	0.8859	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

表 4. 評価

学籍番号	氏名	評価
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．
16344217	津上 祐典	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の理論の方は理解できたが，プログラムの方
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，

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.