

# 知能システム学特論レポート（第hdp2班）

2016年7月14日（中間発表）

## 1 出席者

16344203 井上 聖也

16344216 田中 良道

16344217 津上 祐典

16344229 沈 歩偉

## 2 概要

### 2.1 Hadoop とは

Hadoop とはビッグデータを複数の PC で分散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のによって分散処理を実現している。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。身近な例で言うとクラウドやネットワーク HDD(NAS) などが挙げられる。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。

### 2.2 Spark とは

Apache Spark は UC Berkeley の AMPLab にて開発された大規模データの分散処理フレームワーク。RDD の導入によって, 計算速度が Hadoop より大きく上がっていた。特に再帰など頻

繁的にデータを読むと書くのアルゴリズム，性能が100倍上がれると言われている．しかし性能の代わり，メモリの消耗もHadoopより遥に激しくなる．Spark自身が分散ファイルズシステムを持っていないため，他の分散処理ファイルズシステムと連携を取らなければいけない．HadoopのHDFSは其中でよく使っているシステムの一つ．

## 3 機械学習

### 3.1 データセットとアルゴリズム

学習テーマはスパムメールの分類とした．データセットとしてSpambase Data Setを用いた．このデータセットは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おり，すでに57次元のベクトルとして特徴量が抽出済みである．学習アルゴリズムとして，ロジスティック回帰，ナイーブベイズ，SVM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．正しく分類される確率を最大化（学習）することがこのアルゴリズムの目的である．パラメータを決定する際には確率的勾配法や最急降下法，準ニュートン法などが挙げられる．ナイーブベイズとは，ベイズの定理を用いた分類アルゴリズムであり，各クラスに分類される確率を学習し，最も確率の高いクラスを出力する．SVMとは，訓練データのクラス同士で一番近いサンプル（サポートベクトル）と分離超平面との距離が最大になるように学習を行う線形分離器である．

### 3.2 評価法

学習精度の評価に際して，今回はホールドアウト法を採用しデータセットを7:3の割合で訓練データとテストデータに分けた．ホールドアウト法とはデータをある割合で訓練データ，テストデータに分割し，学習結果を評価する一手法である．他の評価方法として，k-分割交差検証(k-fold Cross Validation)やLeave-One-Out交差検証(LOOCV)などがある．また，訓練データで学習したモデルでテストデータの分類を実行し，そのときの再現率，適合率およびArea Under P-R curveを算出して評価値とした．メール分類における再現率とは，スパム（もしくは

表 1: 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 2: 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

は非スパム)の全体数に対するスパム(もしくは非スパム)と正しく判定できたものの割合である。また適合率とは、スパム(もしくは非スパム)だと分類して本当にスパム(もしくは非スパム)だったものの割合である。スパムメールの分類においては、スパムに対する適合率が低いと非スパムメールを誤って読み過ごすことになるので、これを高く保った上でその他の評価値もなるべく高くするのが良い。また、再現率と適合率は一般的にトレードオフの関係にある。Area Under P-R curve とは、陰性・陽性の判定閾値を変動させたときに再現率と適合率の関係を描写したものである PR 曲線の図における、曲線よりも下の面積を全体の面積の比率で表現した値であり、1に近いほど高い分類性能であることを示している。

## 4 実験

3つのアルゴリズムを用いてメールの分類を行った。その結果を表1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2に示す。

## 5 考察

スパムメール分類においてはロジスティック回帰の性能が高いことがわかった。また，正則化項の設定は分類性能に大きく関わってくるので，ライブラリのデフォルト値を用いるのではなく問題に会わせて個別に設定するのが良い。今回は実験的にこれを求めたが，理論的に求める方法があるのか調査する必要がある。

## 6 今後の展望

分散によって学習時間などにどのような影響があるのか調査する

## 7 進捗状況

- スпамメールのデータセットに対してスタンドアローンモードで分類させた。
- メンバーの PC で Hadoop,Spark の完全分散環境を構築した。
- Spark の完全分散モードで機械学習のプログラムが実行できない。
- 分類後の評価の方法について調査した。

## 8 その他

- 今後，Hadoop,Spark を組み合わせて完全分散処理を行う。
- 学習パラメータを調整してみる。
- 特徴量が抽出されていないデータに対しても分類してみる。