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7月14日

1　テーマ

機械学習の分散並列処理

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 はその中でよく使っているシステムの一つ．

2.3 機械学習

2.3.1 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．

2.3.2 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，正解数に対するスパム (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

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

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．

2.4 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．

3　自分の分担範囲

4　感想

5　評価

16344203 井上 聖也

16344216 田中 良道

16344217 津上 祐典

16344229 沈 歩偉

参考文献

[1] 杉山将, ”イラストで学ぶ機械学習”, 講談社, 2013.

[2] 下田倫大ら, ”詳解 ApacheSpark”, 技術評論社, 2016.

[3] 海野裕也ら, ”オンライン機械学習”, 講談社, 2015.