技術解説

ビッグデータ処理による機械学習・データマイニングの最前線

Frontier of Machine Learning and Data Mining with Big Data Processing

石井　一夫
Ishii Kazuo

インターネットやクラウドコンピューティング，データ処理技術の進歩により，ビッグデータ処理が話題となっている。ビッグデータ処理は，ウェブログ解析やレコメンデーションに始まり，電子カルテやゲノム解析を中心とする医療ビッグデータ，気象ビッグデータ，ビッグデータマーケティングなどいろいろな分野に波及している。その処理エンジンとしての，ディープラーニングなどの機械学習やデータマイニングも注目されている。本稿では，ビッグデータ処理による機械学習，データマイニングについて解説する。

In the progress of the Internet and cloud computing, data processing technologies in big data becomes current trend. The big data processing involves a variety of applications from web log analysis and recommendation to medical big data, such as electronic medical records (EMR) and genome analysis, weather and climate big data and big data marketing. Machine learning and data mining such as deep learning are also popular topics as data processing engines. Here the overviews and future prospects of machine learning and data mining in big data processing will be described.

キーワード：ビッグデータ処理，機械学習，データマイニング，予測分析，オープンソース

1 はじめに

機械学習とは，人工知能における研究課題の一つで，人間が自然に行っている学習能力と同様の機能をコンピュータで実現しようとするものである[1)]。大量のデータを集め従来にない知見を見出すビッグデータの時代では，特にその応用に期待が集まっている。機械学習はインターネット等から情報を取り出す検索エンジン，医療診断，スパム（迷惑）メールの検出，金融市場の予測，DNA配列の分類，音声認識や文字認識などのパターン認識，戦略シミュレーションゲーム，（自ら）学習するロボット，など幅広い分野で用いられている。

一方，データマイニング（マイニングとは「採掘」を意味する）[2)]とは，統計学，パターン認識，人工知能等のデータ解析の技法を大量のデータに網羅的に適用することで知識を取り出す技術である[3)]。通常のデータの扱い方からは想像が及びにくい，ヒューリスティク（heuristic，発見的）な知識獲得が可能であるという。とくにテキスト（文字列情報）を対象とするものをテキストマイニング，そのなかでもインターネットのウェブページを対象にしたものをウェブマイニングと呼ぶ。

両者ともほぼ同様のツールを使い，ビッグデータ処理技術として重複する部分も多い。本稿では，両者を取り分け区別せずデータ分析手法としての概要，およびツールの最新トレンドを紹介し，展望を示す。

2 機械学習の 10 の応用例[4)]

機械学習は，大まかにいって，「分類」と「識別」の２つの機能がある。たとえば，ウェブ上に紹介されていた機械学習の事例を紹介する。

① スパムメールの検知：受信ボックスのなかのメールメッセージの，どれがスパムメールでどれがそうでないかを識別する。
② クレジットカード不正検知：顧客のクレジットカード取引履歴から，それらの取引がその顧客によってなされたものか否かを識別する。
③ 数字認識：封筒の上の手書きの郵便番号が書いてあるとき，その手書き文字の数字を識別する。
④ 会話理解：人の会話の音声パターンから発声している内容の識別を行う。
⑤ 顔検出：デジタル写真の特定の人物が写っている写真を他の人物と区別して識別する。

⑥ 商品レコメンデーション : レコメンデーションとは，過去の購入履歴等から顧客一人一人の趣味や読書傾向を探り出し，それに合致すると思われる商品を，ホームページ上で重点的に顧客一人一人に推奨する機能である。例として Amazon.co.jp では，そのユーザーが過去に購入したり閲覧したりした商品と，類似の商品のリストが自動的に提示されることが挙げられる。

⑦ 医療診断 : 臨床診断において，膨大な臨床検査データと医師の診断結果から相関する検査項目を選択して病気の患者や薬剤効果のある患者を識別することで，医療診断の支援を行う。

⑧ 株式取引：現在と過去の株式の値動きから，株価の変動規則を識別し株価を予測する。

⑨ 顧客セグメンテーション：あるユーザーが試用期間に取った行動のパターンから，すべてのユーザーの過去の行動データをもとに，有料バージョンへ移行するユーザーとしないユーザーを識別する。これらの結果は，マーケティング戦略に利用したり，より見込みの高い顧客にアプローチをしたりする判断を支援する。

⑩ 形状検出：ユーザーがタッチ・スクリーン上に手書きした形について，既知の形状データから，そのユーザーの描こうとした形状を識別する。

3 機械学習法と数理モデリング

機械学習に関連する用語について**表 1** に示す。

3.1 教師付き学習と教師なし学習

機械学習は，おおまかに教師付き学習と教師なし学習に分類される。ひとことでいえば，教師付き学習は「識別」，教師なし学習は「分類」という機能を持っている。詳細を以下に説明する。

（1）教師付き学習

あらかじめ分類などの情報が紐付け（付与）された入力（数値）データが与えられる。たとえば，患者のデータと健常者のデータ，またはスパムメールかそうでないか，などのデータである。これで，入力データをもとに分類するデータのパターンが学習される。学習過程が終了したあと，新たなデータが入力されたときに，これがどのカテゴリに属するかを判定する。

多変量解析の一種である判別分析，サポートベクトルマシン，ベイジアンフィルタリングなどの方法がある。これらの方法で判別関数のような識別関数が得られる。

紐付けデータは，患者か健常者かのような単なるカテゴリー分類だけでなく定量データの場合もある。定量データが出力データであった場合，線形重回帰分析や，ロジスティック回帰分析のような数理モデリング（一般化線形モデル）が実施される。また，最尤（さいゆう）推定法やベイズ推定法など，観測データから確率論的に最適な数理モデルを推定する方法もある。

表 1　アルゴリズムによる機械学習のおおまかな分類とディープラーニング

分類	説明	主な分析方法
教師付き学習	トレーニングデータで学習し，それをもとに新たなデータを「識別」する。 ＊識別のためのモデルを作成するため数理モデリングが実施される。 数理モデリングには， ①一般化線形モデル（重回帰モデル，ロジスティック回帰モデルなど）， ②最尤推定法， ③ベイズ推定法などがある。	判別分析 サポートベクトルマシン ベイジアンフィルタリング
教師なし学習	与えられたデータを相互の類似度で「分類」する。	クラスタリング（階層的クラスタリング，k-means 法，SOM（自己組織化マップ）など） 主成分分析など
ディープラーニング	画像認識，音声認識などに人間の脳（神経回路）の仕組みを模したニューラルネットワークの一種を用いた機械学習	

（2）教師なし学習

教師なし学習では，入力データは，各データの数値とその類似性のみに基づいて分類が行われる。主な分析方法としてデータを分類や分割を行うクラスタリングがあり，階層的クラスタリング，k-means 法，自己組織化マップ（SOM）などのクラスタリングの計算方法（アルゴリズム）が知られている。

3.2 予測分析と数理モデリング

特に，教師付き学習において，学習モデルを数式化して，入力されるデータに基づき特定事象が起こるかどうかを予測することを予測分析と呼び，臨床診断や株価予測など応用範囲が広いことから注目されている。

3.3 ディープラーニング

最近よく聞かれるようになった分析手法で，多層化ニューラルネットワークという分析手法を工夫して（スパース・コンピューティング理論と呼ばれる方法により），識別能力を高性能化したもので，ウェブログ解析などに使われるようになっている。

3.4 ベイジアンフィルタリング

データマイニング法のうち最近注目されているベイジアンフィルタリングについて説明する。一般的に，確率論的事象が次のような数式で表現されるベイズの定理に従う。

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

すなわち，A と B という 2 つの事象があった場合に，それぞれの事象の起こる確率を $P(A)$，$P(B)$ とする。$P(B|A)$ は A が起こった時の B の確率で，$P(A|B)$ は B が起こった時の A の確率である。たとえば，A を「喫煙習慣のある人」と B を「肺がんになった人」と考える。この場合，$P(B|A)$ は「喫煙習慣があって肺がんになった人」の割合，$P(A|B)$ は「肺がんになった人のうち喫煙習慣のあった人」の割合である。ベイズの定理では，$P(B)$ を事前確率と呼び，$P(B|A)$は事後確率と呼ぶ。$P(A|B)$は尤度(ゆうど）と呼ぶ。

ベイズの法則を用いたデータマイニングの応用例にスパムメールフィルターがある。これは，A の事象として「メールを受け取った」という事象を，B の事象として「スパムメールであった」という事象を考え，事後確率 $P(B|A)$ に「メールを受け取った場合に，それがスパムメールである確率」を考える。

4 ビッグデータ処理に対応した機械学習ツール

4.1 機械学習用データ分析ツールの概要

データ分析用のツールは，いろいろなものがある。初心者にとって馴染み深いのは，Excel であるが，扱えるデータ量もデータ分析手法も限定されているので，オススメはできない。商用ソフトとしては SAS，SPSS などの統計ソフトが考えられるが，高価である。オープンソースのソフトウェアでは統計解析ツールである R 言語や Python，機械学習用ツールである Weka などがよく使用される。

ビッグデータ処理における機械学習システムとしては，並列分散処理を用いた Hadoop/MapReduce や，その上で動作する Mahout が，あるいは，さらに高速計算が可能な Spark-MLlib がある。

4.2 ビッグデータ処理用プラットフォーム Hadoop/MapReduce および機械学習ソフト Mahout

データ分析をする場合，処理するデータのサイズが大きくなって数億行のサイズのデータや，テラバイトレベルのファイルを検索するという例も増えてきている。このレベルのサイズを超えるデータをハードディスクに保存する場合，一個のハードディスクに乗り切らないので，ファイルをバラバラに区切り別々のハードディスクやコンピュータに保存することが行われる。これが「分散ファイルシステム」である。また，コンピュータクラスタ上の複数のコンピュータに並列的に仕事をさせ，これを集計することにより大規模計算を可能にする。これが「並列分散処理」である。

このような大規模ビッグデータ処理を可能にした総合フリーソフトが Hadoop で，分散ファイルシステムである HDFS と並列分散処理 MapReduce から構成される（**図 1**）。この Hadoop 上で動作する機械学習システムが Mahout である。これによりビッグデータのレコメンデーションやクラスタリングなどの機械学習が可能になる。

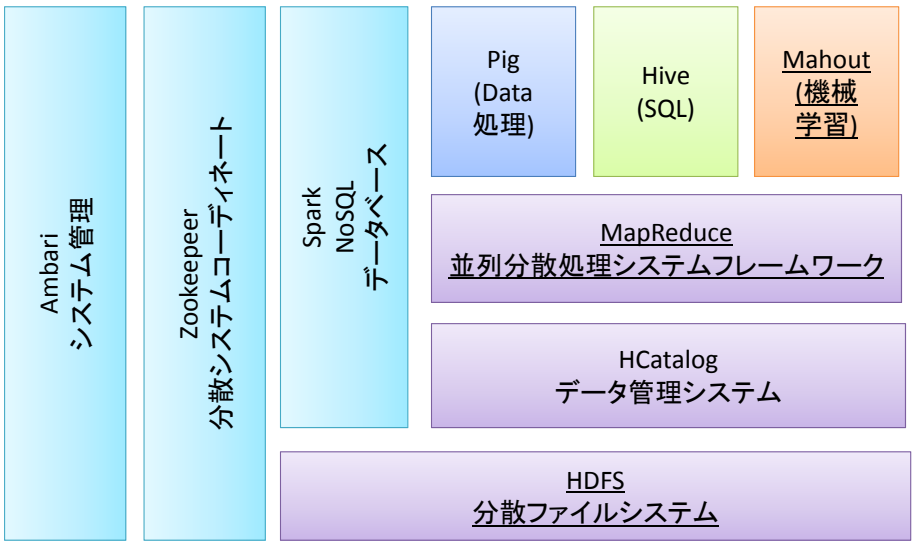

図 1　MapReduce および Mahout が動作するシステムの構成例

4.3　Spark/MLlib

MapReduce は，ハードディスクレベルで分散処理を行う（図 1）。それをメモリレベルで分散して計算処理を行うと，計算の高速化を図れる。これを実装したものが Spark（Apache Spark）である（図 2）。すなわち，Spark は，メモリレベルで並列演算処理を実現するインメモリー並列分散化処理を行うためのプラットフォームである。Spark はメモリレベルでデータを処理するため，ハードディスクレベルでデータを処理する MapReduce より 100 倍速いといわれている。

この Spark で実装された機械学習用ライブラリが MLlib である。MLlib では，サポートベクトルマシン，ロジスティック回帰分析，線形回帰，k-means クラスタリングなどができる。Spark は，2015 年の注目技術として日経 BP 社の IT インフラテクノロジー AWARD 2015 に「準グランプリ」に取り上げられており [5]，現在急速に普及している。

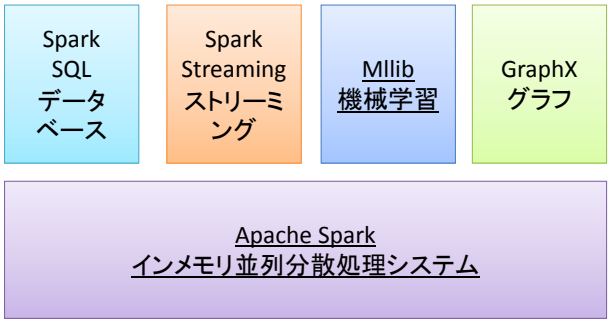

図 2　Spark の構成例

5　終わりに

ビッグデータ処理による機械学習，データマイニングについて，その応用事例に始まり，内容を解説し，最新のツールについて述べた。今まさに現在進化中の技術であり最新情報をできる限り掲載することに努めた。

問題点としては，Spark-MLlib や Mahout などの最近のビッグデータ用機械学習システムは，比較的機能の簡便なレコメンデーションや，クラスクラスタリングなどの教師なし学習に偏っているという点である。予測分析や数理モデリングで重要となる教師付き学習の実装例は，教師なし学習と比べると少ない。ビッグデータ処理用の機械学習ツールはまだ進化途中であり，今後の発展に期待したい。

＜参考文献＞

1）Wikipedia「機械学習」
https://ja.wikipedia.org/wiki/%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92
2）石井一夫：図解よくわかるデータマイニング，日刊工業新聞社，2004 年 12 月
3）Wikipedia「データマイニング」
https://ja.wikipedia.org/wiki/%E6%A9%9F%E6%A2%B0%E5%AD%A6%E7%BF%92
4）日常にある機械学習の応用例
http://postd.cc/practical-machine-learning-problems/
5）IT インフラテクノロジー AWARD 2015［準グランプリ］Apache Spark
http://itpro.nikkeibp.co.jp/atcl/column/15/010800005/010800003/?ST=itarchi

石井　一夫（いしい　かずお）
技術士（生物工学部門）

東京農工大学農学府農学部
特任教授
e-mail：kishii@cc.tuat.ac.jp