ビッグデータとデータマイニング

大黒 智貴，親泊 泰智
Tomoki OKURO，Yasunori SHINPAKU

1　はじめに

通信回線の大容量化や，ネットワークに接続できる端末の多様化，クラウド技術の発展などにより，世界におけるデータ通信の量は増加の一途を辿っている．

ICT の進化や IoT の発達によるセンサ量の増加によって，大量かつ多種多様な様式のデータが収集可能となった．また，収集したビッグデータの活用方法としてデータマイニングがある．データマイニングとは，データ中に潜む項目間の相関関係やパターンなどを探し出す技術であり，近年，企業活動や医療分野などでの応用が期待されている．

2　ビッグデータ

2.1　ビッグデータとは

ビッグデータとは一般的に，画像データや音声データなどの非構造化データを含む大容量かつ多様なデータのことを指す．しかし，明確な定義は定まっていない．ビッグデータの概念図を Fig.1 に示す．

ビッグデータの構成要素は，量，速度，多様性である．量とは，収集したデータ量がペタバイト，ヘクサバイト級であることを指す [1)]．速度とは，常時発生する金融データのように，データの発生頻度が高いことを指す．多様性とは，センサから発生するデータや SNS 上に投稿されたデータなど，データの種類が多岐にわたることを指す．ビッグデータという概念が生まれた背景として，量的観点から見ると，近年，情報機器が扱えるデータ量が急増したことがある．速度的観点からはスマートフォンやモバイル機器の急速な普及による SNS ユーザの増加が挙げられる．多様性の観点からみると，ユーザが音声や動画などを気軽に録音，撮影し，アップロードできる環境になったことが挙げられる．

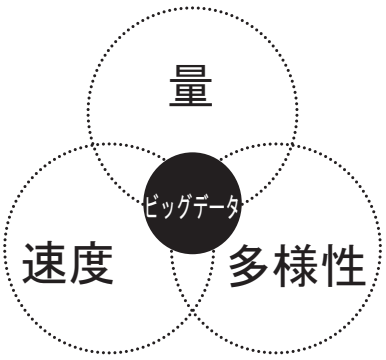

Fig.1　ビッグデータ

2.2　ビッグデータの具体例

ビッグデータの具体例として，顧客のスマートフォンの GPS 位置情報や防犯カメラの映像など大量に蓄積されたセンサデータがある．加えて，SNS 上に投稿されたテキストや音声，動画などのソーシャルメディアデータもビッグデータであると言える．その他，長期的に蓄積された株価や為替レートといった金融データもビッグデータの一つである．

2.3　ビッグデータの活用方法

ビッグデータを活用するためには，データが持つ特徴に応じて類似するデータ同士を分類する処理やデータ間の相関性を求める処理などが必要となる．具体的にどの処理を適用するかは，解決したい問題の種類や収集したデータの性質によって左右される．処理の一つとして，近年注目されているのがデータマイニングである．データマイニングを行うことで新たな知識や，今まで気付けなかった法則性に気付くことができ，企業や行政は，ビジネスや政治に有効利用することができる．つまり，企業や行政がビッグデータを積極的に収集する目的はデータマイニングを行うことにあると言える．

3　データマイニング

3.1　データマイニングとは

データマイニングとは，大量のデータから新たな知識や法則性を導き出すための解析手法である．データマイニングは，テキストや数値だけでなく，音声や動画など，様々なデータに対して行うことができる．データマイニングは，ビッグデータの概念が生まれた 2000 年代より以前に存在したが，近年では主にビッグデータに対して行う解析手法のことをデータマイニングと呼ぶ．

3.2　知識獲得のプロセス

データマイニングにおける知識獲得のプロセスは以下のとおりである．

1. 最適化したい業務上の問題点を理解する．
2. 必要なデータの収集および，各分析手法に適した形にデータを整えるなどのデータの最適化を行う．
3. 収集したデータに適した解析手法を適用し，結果をグラフや樹形図に視覚化する．
4. 知識を獲得する．

以上の流れにより，データから今まで明確でなかった関係性や知識を獲得することができる．

3.3　データマイニングにおける解析手法

データマイニングにおいて主に用いられる解析手法はクラスタリングとクラシフィケーション，アソシエーション

である．

クラスタリングとは互いに類似するデータ同士をまとめる手法である．クラスタリングの概念図を Fig.2 に示す．最初に，任意の数，クラスタを作成し，各クラスタ内の重心を求める．次に，一番近い重心のクラスタにデータを分別しなおす．これを繰り返すことでデータ群を任意のクラスタに分類することが可能となる．具体例として，製品に対するアンケートを自動的に要望，クレーム，故障情報の 3 つに分類するという作業がある．

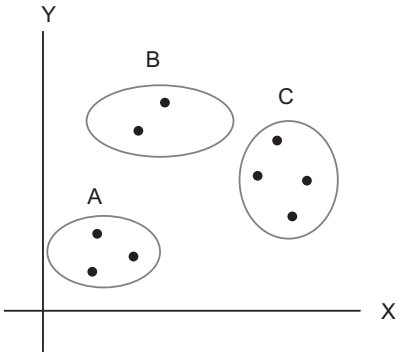

Fig.2　クラスタリング

クラシフィケーションとは既に分類されているデータを利用して，分類条件や分割のルールなどを導き出す手法のことである．条件分岐の木構造である決定木を作成し，未知のデータを逐次分類していくことで，カテゴリを推定する．具体例として，理工学部の卒業論文を履修する必要があるかないかといった例の決定木を Fig.3 に示す．

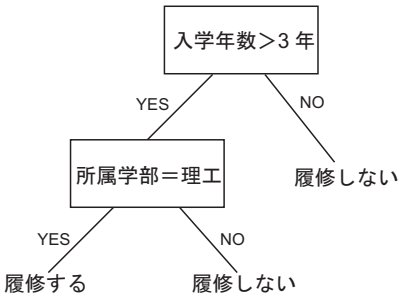

Fig.3　決定木によるクラシフィケーション

アソシエーションとは事象間の関連性を抽出する手法である．具体的には，全データ数中の事象 A と B を共に含むデータ数（支持度），事象 A を含むデータ数中の事象 A と B を共に含むデータ数（信頼度），事象 B の支持度分の信頼度（リフト値）をもとに，関連性を評価する．信頼度と支持度，リフト値が高ければ高いほど，関連性が高いと言える．具体例として，商品 A と B の関連性を調べ高ければ，商品 A と B を並べて陳列することで，売上の向上を図るといったことがある．

4　ビッグデータ分析の問題点

ビッグデータを収集するにあたり問題となるのが，プライバシーである．顧客の個人情報は法律によって保護されているため，企業がユーザデータを収集し，第三者に提供する際に問題になることが多い．

ビッグデータを分析する際に問題となるのが，アナリスト不足の問題である．近年，企業は積極的にビッグデータを用いているが，高度なスキルをもった人材が不足している．加えて，技術の広範囲化という問題もある．これは，医療分野の業務改善であれば，ビッグデータを分析する能力以外にも医学に対する知識が必要というように，技術や知識が広範囲化していくということを指す．

ビッグデータ分析によって得られたデータを活用する際に問題となるのが，国民性の問題である．国によって文化や背景は様々であるので，マーケティングの際に障壁となることが考えられる．

5　今後の展望

自然言語を理解し，学習し，予測する技術であるコグニティブ・コンピューティング分野においてビッグデータ分析の積極利用が行われている．ビッグデータを高速で処理し，適切な返答を行う質疑応答システムである IBM 社の Watson は，米国のクイズ番組でチャンピオンを破り注目を浴びた．今後は，更なる予測精度向上による人工知能分野への貢献が期待できる．

医療分野では，ウェアラブル端末を用いた病気の発病予測が可能になりつつある．ある病気の患者の個人データを大量に集め，ウェアラブル端末によって収集したユーザの血糖値や運動量などが，患者のデータに類似しているかどうかで病気を予測する．ただし，ウェアラブル端末はまだ普及しているとは言えないので，今後の普及と，病気予測技術の実用化に期待したい．

ビッグデータ分析の問題点の一つであるアナリスト不足への対策として，日本政府は官民一体となってアナリストを育成するための政策的な支援を検討している．具体的には，2015 年頃までにビッグデータの利活用などにより約 2 兆規模の市場を創出すること．また，2020 年頃までにビッグデータの利活用などにより約 10 兆規模の市場を創出することが目標とされている．

6　おわりに

IoT の発展により今後流通するデータ量は増える一方であり，情報機器が一度に処理できるデータの量もかつてとは比べものにならないレベルで増加している．これらのデータから新たな知識や法則性を抽出するデータマイニング技術の今後に大いに期待したい．

参考文献

1) Doug Laney, ”3D Data Management Controlling Data Volume Velocity and Variety” ,META Group ,2001.
2) 北林 宏樹, ”身体データの解析による健康情報マイニング” ,DEIM Forum ,2014.
3) 工藤 卓哉, ”データ分析からサービスの改善へ” , アクセンチュア株式会社 ,2015.