IoT eラーニング

データ分析
（分析手法、データマイニング）

国立大学法人　琉球大学

目次

- ビッグデータの分析
  - 大量のデータから有益な情報を探し出す
  - データマイニング
  - 機械学習
  - 統計分析

- ビッグデータ分析手法
  - クロス集計分析
  - アソシエーション分析
  - バスケット分析
  - 因子分析
  - ABC分析
  - クラスタ分析
  - ロジスティック回帰分析
  - 線形回帰分析
  - 主成分分析
  - 決定木分析
  - グレイモデル
  - 独立性の検定
  - 軽量時系列分析

ビッグデータの分析

- **大量のデータから有益な情報を探し出す**

ビッグデータで扱われるデータは、膨大で多様性に富み、整形されておらずノイズが多いデータも対象としている。これらのデータを収集するだけでは価値を見出すことができずデータの塊に過ぎなくなる。

ビッグデータの価値はデータの大きさではなく、その中からビジネスなどに使える有効な情報を求められることにある。

大量でさまざまなデータの中から有益な情報を導き出すには、闇雲に作業を進めるのではなく、分析のための手法を用いる必要がある。

「データマイニング」と呼ばれる手法があり、これは大量のデータに対して統計学や人工知能などの技術を駆使し、データ間の相関関係や隠れた構造などを見出すための手法である。

データマイニングと呼ばれる通り、データから有益な情報を採掘（マイニング）する技術である。

ビッグデータの分析

- **データマイニング**

データマイニングには大量のデータを必要とし、大量にデータがあれば有益な情報を採掘できる可能性も高くなる。

そのために、データを集め保管しておく必要がある。
保管されたデータは改変されず削除されていないことにより、有益な情報を採掘する可能性が高まるので、いわゆる「データベース」に格納するのではなく「データウェアハウス」への格納することとなる。

「データウェアハウス」は文字通り、データの倉庫（ウェアハウス）であり、データを蓄積することを目的としているため、データの削除や更新を原則として必要としない。

データマイニングは、ほとんどの場合、コンピュータを用いて行う。
そのため、非定型で多様性があるデータをシステムに合わせて加工する必要がある。

使用するシステムの分析手法を考慮し「データの形式」を統一するなどの加工を施す。

このようなデータを加工することを「クレンジング」と呼ぶ場合もあり、効率よくデータマイニングを行う上で重要なステップとされている。

ビッグデータの分析

- **機械学習**

データマイニングには、さまざまな手法や技術がある。
代表的なものの一つに「機械学習」による手法がある。

「機械学習」では、事前に仮説を立てる必要がなく、データの中から機械（コンピュータ）が自己学習し相関関係を導き出し、新しい顧客分類などを発掘してくれる。

または、

- 顧客単位での販売手法の選定
- 発生した事象に対する原因の特定

など、人ではできない複雑な条件が絡む問題の分析や最適などを得意としている。

ビッグデータの分析

- **統計分析**

確率統計論などを利用したデータマイニングの手法となる。

代表的な手法として、ロジスティック回帰分析やクラスタ分析、因子分析などがある。

統計分析を行う上で多くの場合、
- 事前に仮説をたてる
- 必要なデータを集める
- 検証したい課題や事象に合わせた分析手法を選択する

といった手順を経てから分析を行う。

分析された結果を検証したうえで、場合によってはこの作業を繰り返し実行する。

これにより有益な情報を発掘していく。

ビッグデータ分析手法

- **クロス集計分析**

設問に対して回答者の性別や年代などの属性項目を交えて集計する分析手法である。
アンケート調査によく利用されている。

例えば、スマートフォンユーザーに「スマートフォンを１日に使用している時間は？」というアンケートを実施し、アンケート結果に性別と年代といった属性を加えることで、縦軸／横軸をモデルとした集計が可能となる。

クロス集計分析を用いることで、性別や地域などそれぞれの傾向をとらえることができるので世論調査などにも利用されている。

- **アソシエーション分析**

アソシエーション分析とは、物事の出来事や事柄などの関連性を導き出すことで、顧客が求めている商品やサービスなどを導く手法である。

有名な例えとして「紙おむつとビール」がある。

あるスーパーで紙おむつとビールが同時に買われることが多くあった。
分析してみたところ、紙おむつの買い出しを頼まれた父親が一緒にビールを購入していたことが分かった。
この結果に基づき、両商品の棚を並べたところ売上の増加につながった。

関連性を導き出すことができれば、ユーザーの行動を掴むことができる。

ビッグデータ分析手法

- **バスケット分析**

基本的には「アソシエーション分析」と変わりはない。
分析対象がユーザーが購入した商品に限定されているところが違う。

ユーザーがバスケット（買い物かご）にどのような商品を入れたかを分析することで、
ユーザーの購買パターンを分析する手法である。

ECサイトを利用すると「この商品を購入した人は、こちらの商品も購入しています」と進めてくるが、これはバスケット分析を利用したものである。

ユーザーが同じ趣向を示していれば、同じ趣向を持つ他のユーザーが購入した商品を購入する可能性があることから、売り上げの増加が見込める分析になる。

ビッグデータ分析手法

- **因子分析**

  さまざまな内容を持つデータの中から共通する因子を見つけ出すことで、データ間の関連性を把握する手法である。
  ビジネスや研究分野など幅広く利用されている手法である。

  膨大でさまざまな内容を持ったデータの中から共通する因子を見つけ、その因果関係を求めることができればデータの相関図を作ることができる。

  こうした分析の結果を参考にして、効率的に売り上げの向上を図る行動を起こすことができる。

ビッグデータ分析手法

- **ABC分析**

在庫管理で多く利用されている分析手法であり、商品や物事に順位付けすることで状況を整理する手法である。

倉庫の中には様々な単価の原材料や製品があり、管理されている個数も違っている。

これらを効率よく管理するために、単純な単価や個数などではなく、個数と単価から求めた金額やよく動くモノなどを基準に順位付けをして管理する手法となる。

また、この手法によるモノの動きをとらえることで購買状況などの情報を知ることができる。

ビッグデータ分析手法

● クラスタ分析

クラスタ分析とは、異なる性質のモノが混ざり合っている集団に対し、互いに似ているものを集めてグループ（クラスター）を作り、対象を分析し相関関係などを導き出す手法である。

客観的な基準と科学的な分類ができることから、マーケティング調査においては、

- ブランドの分類（ポジショニング確認）
- イメージワードの分類
- 生活者の分類

などに用いれられている。

クラスタ分析を用いることで、メーカーサイドの視点ではなく、生活者サイドの視点から見た分類を見つけ出すことができる。

ビッグデータ分析手法

- **ロジスティック回帰分析**

  ロジスティック分析とは「YES or NO」でデータを収集し、物事の発生率を求める分析手法である。

  例えば「商品Aと一緒に購入されることの多いものは何か？」といったデータを求めるのではなく、「商品Aは買われたから、買われなかったのか？」という二者択一で発生率を分析することである。

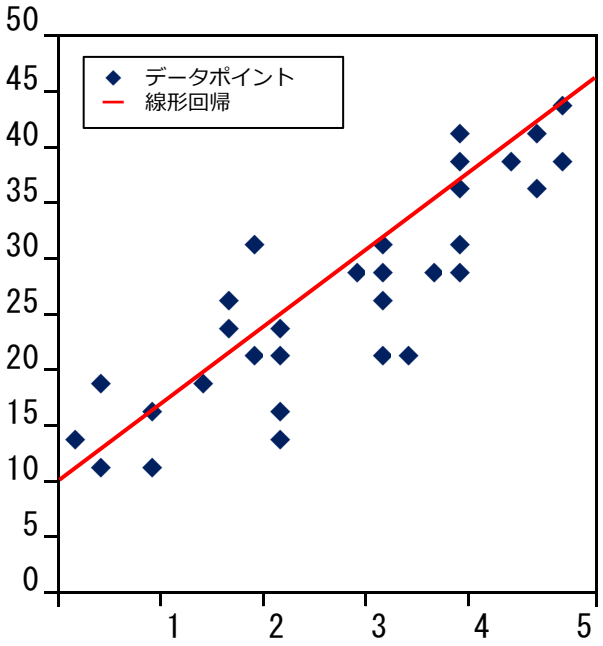
  発生率を予測することができれば、幅広い状況での活用が可能となる。

- **線形回帰分析**

クロス集計分析では、得られたデータをもとにグラフとして分析結果を表すと、各データを通る曲線によって表される。

線形回帰分析では、このクロス集計分析で得たグラフに、論理的に考え出された直線を引くというデータ分析手法になる。（下図を参照）

各データの相関関係を知ることができる。



ビッグデータ分析手法

- **主成分分析**

主成分分析は「次元の縮約」とも言われている分析手法である。

多数の項目を持つデータに対して、少数の項目へ置き換えることによりデータ全体の視認性をよくしてから分析を行う手法となる。

データが持つ項目が多くなるほど複雑化するため、分析しやすい状況を整えてから分析を行う。

一方、項目を縮約するということは「情報の一部を捨てる」ということでもあり注意が必要な手法でもある。

ビッグデータ分析手法

- **決定木分析**

「もしも…だったら」という仮説を立て繰り返すことで結果を予測する手法である。

一つの原因から樹木の枝分かれのようにいくつもの予測を立てていく手法であることから「決定木」と呼ばれている。

リスク管理などで主に利用されており、計画を立案し目標に到達する意思決定を助ける分析手法である。

ビッグデータ分析手法

● グレイモデル

グレイモデルとは、過去のデータをもとに分析し、それに続く数値をグレイ法（灰色理論）で予測する分析手法である。

- 明白なデータは「白」
- 不明なものを「黒」
- 曖昧な状態を「灰色」

と定義する。

白と黒のデータをもとに、灰色のデータ（今後）を予測するデータ分析手法となる。

事象を色で表すことが大きな特徴となっている。

決定木分析同様にリスクマネジメントなどに用いられる。

ビッグデータ分析手法

- **独立性の検定**

  独立性の検定とは、クロス集計分析と共に用いられるデータ分析の手法である。

  ２つのデータの間には関連性（独立性）はあるのかを確認する。

  ２つ以上の分類基準を持つクロス集計表で、分類基準の間に関連があるかどうかを検定することである。

ビッグデータ分析手法

- **軽量時系列分析**

さまざまな時系列データは、一見すると株価のようにランダムに変化しているように見える。

しかし、株価の動きの根底には、景況感や雰囲気、駆け引きが要因にあると言うことができる。

このように時系列データの間で動学的関係を明らかにすることで、ビジネスやマーケティングにおける理論や仮設などを検証するための分析手法である。

時系列で変動するデータの流れを解析し、それをもとにマーケティングを展開していくという事である。