1章　データマイニングとは（1）

10NM706F　江口晃

統計学

- 統計学の利用
  - 医学
    新薬の薬効検定
  - 経済学
    景気動向の予測
  - 政治・社会分野
    選挙速報

その他にも生物学、農学、商学、心理学、言語学…

1.1　統計学の憂鬱

- 統計学　→　記述統計学と推測統計学
  しかし推測統計学は大きなダメージを受けている。

  – **検定論**
    理論的枠組みは1950年代に既に完成している。
    理論的枠組みに欠陥がある。
  – **分布論**
    複雑で煩雑な公式で実用されない。

1.2 データマイニング

- データマイニングとは

  定義:

  「有用で、かつ既知でない知識をデータから抽出する自明でない一連の手続き」

  実状:

  大量のデータを分析するのに有効な手法の集合体。

  ビジネスデータに対して現実的な要請により発展させてきた技術体系である。

1.3 データ解析の新しい流れ

- **非線形性**

  特定の関数形に限定することなく変数間の関数関係を発見できる。

- **視覚化**

  データの特徴を直感的に分析。ビジネスデータ解析に有効。

- **交差妥当性**

  母数の推定に利用しなかったデータでモデルを評価する。

- **最適性・一意性のなさ**

  最適な解に到達できる保障はない。結果が一定していない。

1.4 データウェアハウス

- データウェアハウス

  継続的に収集される大量のデータを分析しやすい形式で運用するための考え方。

- POS(point of sales)

  販売時点の情報。バーコードで商品、価格、販売個数などが記録される。

1.4.1　基本的性質

- **時間的依存性**

  長期に渡ってデータを収集することが望ましい。
- **不揮発性**

  データの破棄や書き換えをしない。
- **サブジェクト指向**

  サブジェクトごと(売上げ、商品など)に蓄積する。
- **統合**

  単位を統一し、定められたフォーマットでデータを蓄積する。

1.4.2 データの準備（1）

- **データの前処理**

  必要な部分をコピーし、目的に応じて編集したデータマートを用意する。

  データマイニングに要する時間の90%はデータの準備に費やされる。

1.4.2 データの準備（2）

- データの選択

  どの変数を利用するかを決定する。

  基準変数や予測変数にどれを選ぶかなど。

- レコードの再集計

  分析目的に合わせて適切なレコードに再集計してから分析する。

  時間・空間・対象などの単位を再編集してレコードの数を減らす。

1.4.2 データの準備（3）

- **データの洗浄**

  外れ値、欠損値、不整合なデータを除去・修正する。

- **データの補強**

  別のデータマートからデータを調達し、分析中のデータマートに変数を加える。

  新たな変数を付加してデータの補強をする。

- **データのコード化**

  分析目的に合わせて変数の表現を変える。

  漢字の曜日を実数をとる変数にするなど。

1.4.3 再びデータマイニングとは

- データマイニングの作業

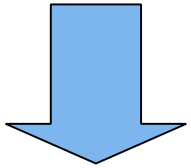
  データウェアハウス内の変数の性質と、それに合うモデルを探し当てるまでに長い時間を要する。

  市場の構造は少しずつ変化し、モデルが合わなくなっていく。

データマイニングは地味な粘り強い作業の連続である。

1.5 データマイニングと統計解析環境（1）

データマイニングでは、大規模なデータ処理に対応した計算機とソフトウェアが必要。



最近ではパソコンとデータ、統計知識さえあればどこでもデータマイニングできる。

1.5 データマイニングと統計解析環境(2)

- 統計解析環境R
  - 統計用途に特化されている。
  - オープンソースである。
  - 世界中の統計学者がR用の統計パッケージを無償で開発・提供している。

1.6 コピーレフト、オープンソース、そしてR

- コピーレフト

  ①プログラムを作成したらソースコードを公開する。

  ②ソースコードの著作権は破棄しない。

  ③自由に使用・改良・再配布可能。

  ④改良したソースコードには①～③の条件が適用される。

  Rはコピーレフトの下に利用が許可されているソフトウェアであり、発展は商用統計ソフトより圧倒的に早い。