データマイニング

松本 大樹，松下 昌平
Taiki Matsumoto，Shouhei Matsushita

1　はじめに

近年，記憶装置の低価格化やネットワーク環境の整備により，社会では大量のデータが生まれるようになった．また，CPU の高性能化や，情報処理技術が向上により，膨大なデータを扱えるようになった．それにより，従来は破棄されていた大量のデータからユーザに合った有用な情報を取り出す技術として「データマイニング」という技術が生まれ，あらゆる分野での活躍が期待されている [1]．

2　データマイニング

2.1　データマイニングの概要

データマイニングとは，大量のデータの中から規則性，および関連性など意味のあるパターンを見つけ，価値のある情報や知識を発見するデータ処理技術である [2]．ここでいう知識とは，データ間の相関関係やパターンである．データマイニングはあらゆる分野でが期待されており，現在では，ネットショッピング，天気予報，および医療などでも利用されている．

データマイニングのイメージ図を Fig. 1 に示します．

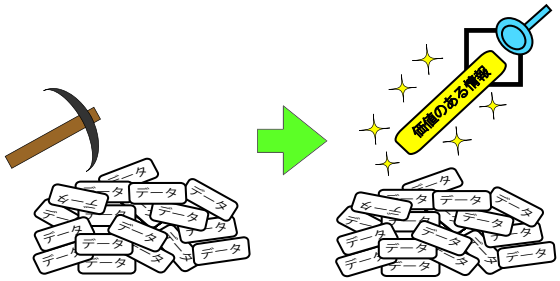

Fig.1　参考画像 1

2.2　プロセス

データマイニングを行う際には前準備が必要である．まず，問題の定義である．どのような問題があり，どんな情報を獲得するか，またそれによりどのような解析手法が必要であるかを明確にする．

次にデータを準備する．課題に対してどのようなデータを用いればよいか，あらかじめ吟味しておく．そして，解析を行いやすくするため，データの単位をそろえたり，データが欠落していたら平均値を採ってデータを補うなど，データを解析しやすい形にする．ここまでが前準備である．

その後，データマイニングを行い，視覚化する．視覚化とは，データマイニングによって得た結果をグラフや樹形図などで表し，結果を見やすくすることにより，他の人に理解できるようにすることである．そしてそれが新しい知識として得られる．

1. 問題の定義
2. データの準備
3. データマイニングと視覚化
4. 知識獲得

3　データマイニングの解析手法

解析対象によっては，一つの手法だけでなく，いくつかの手法を適所に用いる場合もある．そのため，ユーザーの目的に対応したアルゴリズムで解析する必要がある．よく用いられる手法を以下に述べる．

- クラスター解析
  クラスター解析とは，多数の変数を持つデータ群を，似た者同士でグループ分けする手法である．そのデータ群をクラスタと呼ぶ．企業にとって消費者のニーズを知ることは非常に重要なことである．しかしどのような消費者が存在するかを知ることができれば，その打開策が見いだせる．そのような場合によく用いられる有効な手法がクラスター解析である [3]．
  クラスター解析には階層的クラスター解析と，非階層的クラスター解析がある．
  階層的クラスター解析は樹形図を描くことを目的としたもので，データ同士の距離が近いものから併合していき，最終的に一つのデータ群にすることである．
  非階層的クラスター解析は，解析者があらかじめ指定したクラスター数で観測対象を分類する [3]．クラスター数を指定しなければ解析できない非階層クラスター解析は，それなしでできる階層的クラスター解析よりも汎用性が低い．しかし，実際にデータマイニングを行う際には，クラスター数の数に目星をつけ，非階層的クラスター解析を複数回試行したほうがほうが効率が良い場合も存在する．また，大量の項目の分析に用いても結果が安定している．クラスター解析のイメージ図を以下に示す．
  解析するデータ群は Fig. 2 を参照します．
  階層的クラスター解析のイメージは Fig. 2 を参照します．
  非階層的クラスター解析のイメージは Fig. 3 を参照します．
- 回帰分析
  データマイニングにおける回帰分析とは，予測の対

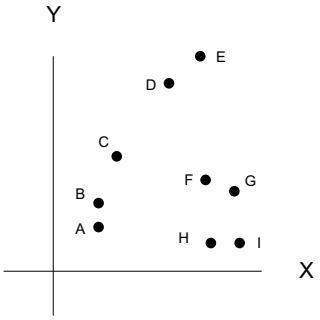

Fig.2　参考画像 2

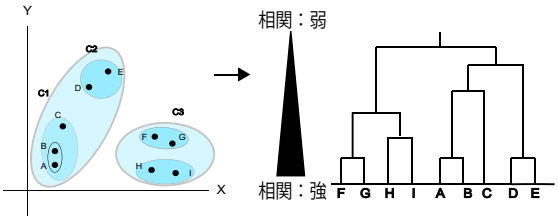

Fig.3　参考画像 3

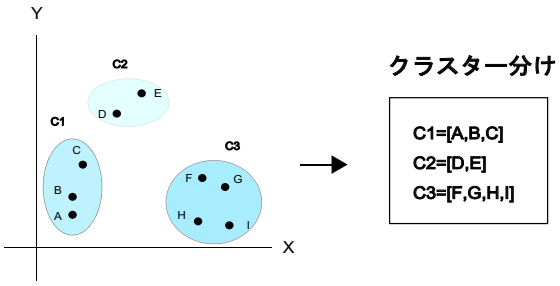

Fig.4　参考画像 4

象となる変数（目的変数と呼ぶ）と，その変数に関係すると思われる観測可能いくつかのの変数（説明変数と呼ぶ）の関係について，過去のデータを用いいて関連性を定式化し，新たなデータに対する目的変数の予測に役立てることである [4]．

- 決定木

決定木とは，データマイニングの中では最もよく利用される解析方法の一つである．大量のデータに対して何らかの基準に基づいて分類を繰り返し，予測を目的とした方法である [?]．分析の方法はシンプルだが，正確な方法である．分類分けの基準は，あるカテゴリーに属するか属さないか，データがある値より大きいか小さいか，などがあげられる．例えば，企業が顧客にダイレクトメールを送りたいとする．決定木を用いて分類すれば顧客データから，どの年齢層，性別の顧客が今までの総購入金額が多いかがわかる．このように多数の変数を持つデータを分類でき，視覚的に理解することが出来る．

4　データマイニングと統計解析との比較

3 節で述べた手法は，統計解析の分野でも用いられている．そこで，本節では，統計解析とデータマイニングを比較しする．

データマイニングと統計解析との大きな違いは扱えるデータ量の違いである．データマイニングは複数のコンピューターを用いて解析するため，数千，数万ものデータを扱うことが可能である．

次にデータの種類が多いことである．データは数値データだけでなく，テキスト，画像，音声データなどさまざまなものを扱う．特に，テキスト文章を扱うものをテキストマイニングと呼ばれている．

また，統計解析は解析者が仮説を立て，それに対して解析を行う仮説の検証である．そのため，仮説を立てることに専門的知識が必要である．また，一つのデータ群からは一つの分析結果しか得られない．しかし，データマイニングは解析手法によって一つのデータ群から同じ結果が得られるとは限らない．仮説検定と仮説発見，少量データからの推定と大量データによる信頼性の保障が統計的データ解析とデータマイニングの違いである [5]．

5　まとめ

Amazon や楽天市場などのインターネットショッピングサイトなどで，サイトに顧客情報を登録しておけば，購買履歴やおすすめの商品を紹介してくれるようなシステムがある．サイトから集められた顧客購買情報はデータベースに集められ，データマイニングにより購買傾向を分析し，オススメ商品を表示する．また，その他にも天気予報や競馬予想，医療など多くの場所でデータマイニングが利用されている [1]．

このように，データマイニングは多方面で活躍しているが，重要なのはユーザがきちんとした目的を持って，それに適した解析方法を選択することである．また，どんなに新しい知識が発見されたとしても，それを扱うのは人間であるので，どのように利用し，今後発展させていくかが重要である．

参考文献

1) 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, 2004-12-20”
2) Usama Fayyad and Gregory Piatetsky-Shapiro and Padhraic Smyth, ”Knowledge Discovery and Data Mining: Towards a Unifying Framewor”,AAAI ,1996
3) 豊田　秀樹, “データマイニング入門　R で学ぶ最新データ解析”, 東京図書,2008-12-25
4) 水田　正弘　山本　義郎　南　弘征, “S-PLUS によるデータマイニング入門”, 森北出版株式会,2005
5) 松田 芳雄,”顧客分析とデータマイニングの動向”,UNISYS TECHNOLOGY REVIEW 第 68 号,2001