ま　え　が　き

本書の目的は，初めてデータマイニングを学ぶ者に対して，データマイニング手法の原理を説明することである。データマイニングは，膨大なデータから半自動的に知識を発見する手法の総称である。従来のデータ分析の手法としては統計がある。統計においては，おもに人間が「仮説」を生成し，その仮説を検証するためにデータを捜査する。一方，おもに計算機がデータを自動的に分析する機械学習という分野もある。ここでは人間は計算機に対してデータの中をどのように捜査するのかのみを指定して，あとは計算機自身が自動でデータを捜査して知識を発見する。発見されたこの知識は計算機自身の中に蓄えられ，人間にフィードバックされるとは限らない。

データマイニングはこの統計と機械学習の二つの分野の間に位置するということができる。データマイニングでは，人と計算機は協同してデータを分析する。例えば，統計では人がおもに役割を担っていた仮説生成を，データマイニングでは時には計算機が担当する。また機械学習では計算機が得た知識を人にフィードバックしていなかったのに対して，データマイニングは人に知識をフィードバックし，そして人はその得た知識を基にさらに深くデータを分析する。

このようなデータマイニングの「半自動」によるデータ分析をよりよく行うためには，データに対して適用するデータマイニングアルゴリズムの原理を理解している必要がある。例えば，仮説というデータを分析する目的が明確にある統計では，データを分析する第一の目的は，仮説が支持されるかどうかを yes/no でわかることである。これに対してデータマイニングの目的は少し曖昧（あいまい）である。分析の当初では，統計学の仮説に相当する明確な目的を持っていないことがあ

る。人はデータを分析しながらデータに対する知識を高めていき，そして何か法則のようなものを発見することを試みる。そのようなデータマイニングでは，データマイニングアルゴリズムの発見する知識は，あくまでもデータを理解するための中間的な情報であると考えられる。そしてその情報からデータ全体の理解を深めていくためには，なぜデータマイニングアルゴリズムがそのような情報を出力したのかを理解することが重要である。そのためには，データマイニングアルゴリズムの原理を理解している必要がある。

本書では，このように人間と計算機が協同でデータ分析することを可能とするために，データマイニングアルゴリズムの原理を説明する。またそのための準備としてデータ形式やデータ前処理などを説明し，そしてデータマイニング手法の評価方法，さらにデータマイニングを適用したさまざまな活用例について示す。

本書の作成に当たり，多くの方々にご指導をいただきました。まず著者の博士後期課程の指導教官であり本書を作成するチャンスを与えてくださった北陸先端科学技術大学院大学 副学長/知識科学研究科 教授 國藤進先生に感謝いたします。また，北陸先端科学技術大学院大学 ライフスタイルデザイン研究センター 准教授 金井秀明先生には，大学院在学中および研究員時代に多大なるご指導をいただきました。本書で取り扱うデータマイニングの適用例の一部は，金井秀明先生のご指導の下に行った研究が基礎となっています。そして本書は，新潟国際情報大学の「知識情報処理」の講義資料をまとめたものです。新潟国際情報大学の教職員および学生の方々の意見が反映されています。この場をもってお礼を申し上げます。

2013 年 2 月

中田 豊久

目　　次

1. データマイニングとその周辺

2. データの形式，事前処理，俯瞰

3. データマイニングの手法

4.　データマイニング手法の評価

5. データマイニングの実践例

1 データマイニングとその周辺

データマイニングとは，膨大なデータの中から有益な情報を発掘（マイニング）する技術の総称である。本章では，このデータマイニングの概要を示し，また他の分野との関係について紹介する。

1.1 データマイニング

データマイニング (data mining) とは，膨大なデータの中から有益な情報を発掘（マイニング）する技術の総称である。取り扱うデータはおもに，伝統的なデータマイニングでは比率尺度の**数値型データ** (numeric data) と名義尺度の**名詞型データ** (nominal data) のみを扱うものが多かった。近年では，テキストデータやネットワークデータなどのさまざまなデータに対するデータマイニング技術が開発されている。それらのデータに特化したデータマイニングをそれぞれテキストマイニングやグラフマイニングと呼ぶこともある。

比率尺度のデータとは，実数で表されたデータのことである。例えば二つのデータを持ってきて，その加算，減算，乗算，除算を行うことができ，その演算結果もまた比率尺度のデータとして利用することができるデータである。一方，名義尺度のデータとは，「田中」，「鈴木」のような名前であったり，数値であっても背番号のように加算，減算等の処理ができないデータのことである。この名義尺度において可能な処理は，同じデータであるか違うかを判定すること，それを基にしてどのデータがいくつあるのかを示す頻度分布と呼ばれる処理などだけである。

表 *1.1* は，ある人がゴルフに行ったか行かなかったかを示す 14 個のデータである。表の列は，ある日の天気，温度，湿度，風，そしてゴルフに行ったか，行かなかったかを○，×で表している。このデータを用いて伝統的なデータマイニングの例について示す。

表 *1.1* ゴルフに行く/行かないのデータ

天気	温度	湿度	風	ゴルフ
晴	暑	高	無	×
晴	暑	高	有	×
曇	暑	高	無	○
雨	暖	高	無	○
雨	涼	普通	無	○
雨	涼	普通	有	×
曇	涼	普通	有	○
晴	暖	高	無	×
晴	涼	普通	無	○
雨	暖	普通	無	○
晴	暖	普通	有	○
曇	暖	高	有	○
曇	暑	普通	無	○
雨	暖	高	有	×

「ゴルフ」：○行った，×行かなかった

1.1.1 ルール指向マイニングの例

ルール指向マイニング (rule based mining) とは，データに潜むデータ項目間の関係を発見する手法である。ルール指向マイニングの例として，ID 3 というデータマイニングアルゴリズムを表 *1.1* のデータに適用する。その結果として**図 *1.1* のディシジョンツリー** (decision tree) と呼ばれるものが自動的に生成される。ディシジョンツリーとは枝と葉で構成された木構造で，枝が分かれる部分に質問があり，その回答の数だけ枝分かれをしている。木の根元の部分からスタートし，質問に回答していくと，ある一つの回答にたどり着くというものである。

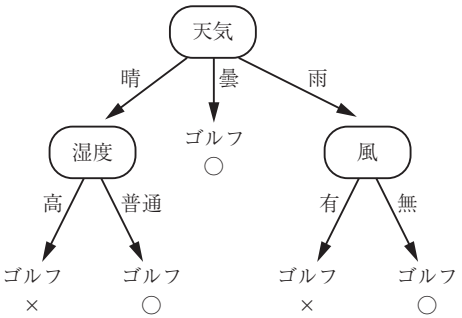

図 *1.1* 表 *1.1* から自動生成したディシジョンツリーの例

図 *1.1* では，枝の質問はゴルフに行く/行かない以外のデータであり，回答はゴルフに行く/行かないである。木の最上位の根の部分ではまず，天気について質問される。そしてその回答が晴れの場合にはさらに湿度について質問され，その回答によってゴルフに行くか行かないかを判定することができる。

1.1.2 相関ルールマイニングの例

ルール指向マイニングでは「ゴルフに行く/行かない」を判定するためのルールを抽出した。ここでは「ゴルフに行く/行かない」というデータ項目に着目するのではなく，すべてのデータ項目を平等に見て，データに潜むルールを抽出する方法を示す。その例として**相関ルールマイニング** (association rule mining) を紹介する。

表 *1.1* のデータに**アプリオリ** (a priori) というデータマイニングアルゴリズムを適用する。その結果として**図 *1.2*** のルールが自動的に出力される。この図に示したルールは，相関ルールと呼ばれる。

ルール 1	if	天気＝曇	(4)	then	ゴルフ＝○	(4)		
ルール 2	if	気温＝涼	(4)	then	湿度＝普通	(4)		
ルール 3	if	湿度＝普通	and	風＝無	(4)	then	ゴルフ＝○	(4)

図 *1.2* 表 *1.1* から自動生成した相関ルールの例

相関ルールの括弧内の数値は，その条件に一致するデータ数を示す。例えばルール 1 は，“天気 = 曇” のデータが四つあり，そのうち四つが “ゴルフ = ○” であると読むことができる。

ルール 1 の「天気が曇りならば，ゴルフに行く」は，ルール指向マイニングの結果のディシジョンツリーの一部と同じである。一方，ルール 2 の「気温が涼しいならば，湿度は普通」には，ゴルフに行く/行かないを推定しようとするルール指向マイニングにはないデータの特徴が現れている。そしてルール 3 の「湿度が普通でありかつ風がなければ，ゴルフに行く」は，ゴルフに行く/行かないを判定しているのであるが，ディシジョンツリーとは少し異なるルールを抽出している。この違いは，アルゴリズムがどのようにルールを抽出するのかの違いによる。

1.2 テキストマイニング，グラフマイニング，Web マイニング

データマイニングの中には，取り扱うデータに特化した名称を持つものがある。取り扱うデータがテキスト情報である場合のマイニング手法のことを，**テキストマイニング** (text mining) と呼ぶ。データマイニングのアルゴリズムにおいても名義尺度としてテキストを扱うことはできる。ここでいうテキストマイニングとは，さらに言語としての情報を扱うものである。例えば，形容詞や名詞といった品詞構造に着目して，意味的な側面からマイニングを試みるような手法である。

グラフデータ（ネットワークデータ）を対象としたデータマイニング手法は，**グラフマイニング** (graph mining) と呼ぶ。グラフマイニングのおもな手法は，グラフの中に頻出する部分グラフを抽出する手法である。

また，Web に特化したマイニング手法を **Web マイニング** (web mining) と呼ぶ。Web マイニングには，HTML ドキュメントの文章からマイニングする Web コンテンツマイニング，ハイパーリンクによる Web のネットワーク構造からマイニングする Web 構造マイニングなどがある。前者はテキストマイニ

ングと，後者はグラフマイニングと親和性が高い。

1.3 統計，機械学習とデータマイニング

データマイニングの手法は，統計や機械学習の手法を基にしているものが多い。本節では，これらの統計や機械学習とデータマイニングの違いについて述べる。

統計においては，人間が仮説を生成し，その仮説を検証するためにデータを処理する。一方，機械学習においては，人間が学び方のみを計算機に伝え，計算機は入力データと出力データを与えられて，その二つをつなぐ方法を計算機自らが発見する。これらに対してデータマイニングは，その中間的な位置にある。データの扱いを統計が「手動」でしているとするならば，機械学習は「自動」でデータを処理することになる。その場合にはデータマイニングは「半自動」となる。「半自動」とは，自動的に発見する部分もあれば，それを人の手で検証したり，やり直したりすることもあるということである。このようなイメージを**図 *1.3*** に示す。

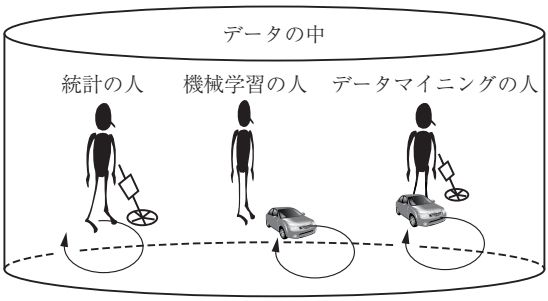

図 *1.3* 統計，機械学習，データマイニングの関係

図 *1.3* は，データの中で何かしらの法則を発見しようとしている三人の人を表している。一人は統計，もう一人は機械学習，そして最後の人はデータマイニングの法則を発見するための支援技術として使用している。統計は，探知機

を持ってデータの中を自分自身が歩いているようなものである。データのどの部分を調査するのかは，人間が決める。一方，機械学習は，法則を発見するためのプログラムを自動車のような機械に載せて自動運転させるようなものである。帰ってきた車には法則が自動的に入っていることになる。そしてデータマイニングは，この自動車に自分自身が乗るようなものである。自動車はいろいろな箇所を自動的に動き回るのであるが，時には人が車を止めてそして車を降りてデータの周りをよく観察する。そして車に乗ってまた新しい箇所を調査する，というようなイメージである。

1.4　発想支援とデータマイニング

発想支援とは，人の創造的問題解決プロセスを支援することである。その発想支援は，発散的思考支援，収束的思考支援などに分類することができる。発散的思考支援とは，ある問題に対して関連する情報や知識をできるだけ幅広く大量に想起抽出する思考作業である発散的思考を支援することである。収束的思考支援とは，発散的思考により集めたバラバラのアイディアの集まりをまとまりのあるものに集約していく思考を支援するものである。データマイニングはその使い方により，この両者に対して適用が可能である。

発散的思考には，概念空間の空間自体の拡張や変容と，概念空間の探索がある。前者は，例えば考えている領域を拡張してより多くのことを思考の対象としていくことである。一方，後者は，領域の中を隅々（すみずみ）まで検討し，例えばとり得る値のすべての組み合わせを検討するような思考である。この後者の探索においては計算機が支援できる要素が大きい。領域における探索は，人が行おうとした場合にはすべてを隈（くま）なく捜すことは困難である。そこで計算機によって網羅的に可能性を調査することによって，自分自身が気づかなかった解にたどり着くことがあると考えられる。

一方，収束的思考では，数多く集まった情報をまとめる作業になる。数多くの情報から何かしらのルールを抽出するデータマイニングは，その収束的思考

索　　引

—— 著 者 略 歴 ——

1993 年　東京工科大学機械制御工学科卒業
1993 ～
1995 年　日本電気ロボットエンジニアリング株式会社勤務
1996 ～
1997 年　株式会社日本総研テクノス勤務
1997 ～
2000 年　株式会社ソリトンシステムズ勤務
2002 年　北陸先端科学技術大学院大学知識科学研究科博士前期課程修了
2003 ～
2004 年　株式会社本田技術研究所勤務
2006 年　北陸先端科学技術大学院大学知識科学研究科博士後期課程修了
博士（知識科学）
2006 ～
2008 年　北陸先端科学技術大学院大学産学官連携研究員
2008 年　新潟国際情報大学講師
現在に至る

基礎から学ぶデータマイニング

Basic Lesson of Data Mining

2013 年 4 月 26 日　初版第 1 刷発行　★

検印省略

著　者　中田豊久（なかだとよひさ）
発 行 者　株式会社　コロナ社
代 表 者　牛来真也
印 刷 所　三美印刷株式会社

112–0011　東京都文京区千石 4–46–10
発行所　株式会社　コロナ社
CORONA PUBLISHING CO., LTD.
Tokyo Japan
振替 00140–8–14844・電話(03)3941–3131(代)
ホームページ http://www.coronasha.co.jp

ISBN 978–4–339–02470–8　（柏原）　（製本：愛千製本所）
Printed in Japan