

データマイニング活用ガイド

概念から実践まで 「改裝版」

P・キャベナ / P・ハジニアン / R・スタッドラー / J・ベルフィース / A・ザナシー 著

日本アイ・ピー・エム株式会社 河村佳洋 / 福田剛志 監訳

日本アイ・ピー・エム株式会社 ナショナル・ランゲージ・サポート 訳



一九五七年頃のことであるが、「私はこの国をくまなく旅して、トップクラスの経営者達と話し合ってきた。データ処理は一時的な流行にすぎない、今年一年もたないと断言できる」と言下に述べて、データ処理の出版提案を拒否した編集者へ、この本を捧げたい。

序文

データマイニングは、大きなデータベースからの情報抽出という問題を取り扱ったため、機械学習、パターン認識、統計、データベース、および視覚化など、種々の手法を合わせた総合的な手法です。この手法が出現したのは、簡単な統計手法と経営戦略支援システムを組み合わせた従来の意思決定支援である限り、今日の激変するビジネス環境で求められる時間内に、現代の大きなデータベースおよびデータウェアハウスを取り扱えないということが理解されてきたからです。データマイニングは、八〇年代半ばまでは特殊な研究分野であったものが、経営と学問の世界の想像力を融和させて、あつと言う間に現代の華々しい分野になりました。事実、「フォーチュン」(Fortune)誌のランキングで上位五〇〇社のうちの

八〇%が、データマイニングのパイロット・プロジェクトを進めているか、あるいは一つ以上のデータマイニング・システムをすでに配備し活用しています。

「データマイニング」は、ユーザーの介入がほとんどなくても、データから情報を自動抽出できる発見型手法を意味していますが、この言葉があまりにもよく知られるようになったため、照会とレポート作成、オンライン分析処理(OLAP)、および統計解析など、他のタイプのデータ解析手法を指すようにもなりました。これらの方法は、厳密には同義語ではありませんが、この発見型手法を補うものであり、現実のアプリケーションのほとんどでは、互いに協力し合うものとして使用されています。とくに、これらの方法は、あらかじめ仮定されたビジネス上の知識を検証するため使用されます。発見型手法は、新しいビジネス知識を識別したり、従来手法により以前に妥当とされた知識をさらに精緻化するのに使われます。本書に載せられているデータマイニング手法の徹底的な習得は、これらを理解する上で読者の大きな助けになるでしょう。

データマイニングは、データウェアハウジングに類似して、アプリケーションによって実行されます。そうし

たアプリケーションのほとんどは、行動を理解すること、もつと具体的に言えば、顧客の行動を理解することを目的としています。つまり、顧客の収益性を獲得し、維持し、生涯的価値を増やす方法を知ろうとするものです。さらに、データマイニング・アプリケーションは、破産や詐欺など種々のタイプのリスクを高めるトランザクションに顧客がかかわっているかどうか、企業が理解するのに役立ちます。そうしたアプリケーションをうまく実現するため、データマイニング手法が果たす役割を理解することは極めて重要です。本書では、ビジネス・コンサルタントが担当した一連の事例研究を用いて、データマイニング手法の役割が解説されています。

データマイニングが出現して、早々に実用化されるにいたった基本的要因は、大きなデータ・セットが入手可能であったことです。企業は、日常的に取り込み保管しているデータのうち、比較的わずかな割合のデータしか分析していないことを認識し始めています。しかし、大きなデータ・セットがあるというだけの理由でデータマイニングを行うと、通常は失敗に終わります。そのため、適用領域の設定とプロセスの規定が必要です。ここに、適用領域はプロセスを順次実行するための全体的な処理

範囲の設定を意味します。また、プロセス（本書で詳しく説明されている）は次のものを規定します。

- 一、マイニングを行う必要のあるデータを判別する方法、すなわち適切なデータ・ソースの識別方法
- 二、選択されたデータの種類のごとに、表現形式を決定する方法、すなわちどのデータを変換して使用するか、そしてどのデータをそのままの形式で使用するか決定するための方法
- 三、データの前処理を行う手法およびデータマイニング用にどの手法を選択するか
- 四、分析目標を達成するため、データマイニングに加えて、どのような補助ツールが必要か
- 五、マイニング処理結果を解釈する方法とその後で意思決定に用いる情報を選択する方法

適用領域を注意深く識別し、系統的に体系化された知識発見プロセスに一貫して従うならば、本書の著者らが確認してきたように、プロジェクトは必ず成功することでしょう。

データマイニング・システムは、単一構成要素のツールから、データベース管理システムにゆるく結合された複数構成要素から成るツールキットへ急速に発展しまし

た。次世代のシステムは、データベース管理システムにしつかり統合されたものになり、データベース内のデータをマイニングすることができるようになります。しかし、そうしたシステムがどれほど洗練されても、そのシステムを使用する企業が自動的に成功するわけではありません。そうした手法を適用する準備ができているかどうかを評価する必要があります。

第一に、企業は自分のデータの品質を確立する必要があります。データマイニング手法は、ある種のノイズは許容するかもしれませんが、ノイズが過剰にあると、マイニングされた情報の品質に影響を及ぼし、場合によっては意味のあるマイニングが行えなくなることをさえます。

第二に、企業は適切な手法をいつどのように適用するかを確立する方法、そしてどうしたらマイニングされた情報を最大限に利用できるか、その判別方法を開発する必要があります。たとえば、現代の企業は、データマイニング手法を適用してミクロ市場を識別することで、販売量を増やそうとしています。そうした企業は、クラスターリング手法によって数百、ときには数千ものセグメント（それぞれがミクロ市場の可能性を持っていると考え

られる）を自動的に識別できることに夢中になってしまい、そうした多くの潜在的市場において企業が今置かれている環境を考慮した上で、実際にうまく販売活動を行うことができるかどうか、そこまでは考えようとしてないようです。実際、適切なデータマイニング・ツールの選択は、この評価の過程における最初のステップではなく、最後のステップでなければなりません。

データマイニング方法の適用が成功したかどうか、どうしたら分かるでしょうか。これは難しい質問であり、その答えはしばしば領域または企業（あるいはその両方）によって異なってきました。企業によっては、投資利益率を基準として成功したかどうかを評価しています。たとえば、ダイレクト・メールを送送する企業は、データマイニング方法でターゲットとした顧客の反応率と特定のキャンペーンにかかった全コストに基づいて、投資利益率を評価しているかもしれません。企業によっては、そうした「しつかりした」基準がありません。そのような場合には、企業が毎日の仕事でデータマイニングを用いる度合いを調べます。それらの企業において、もっとも成功している企業は、データマイニングを電子メールのように使用しています。ビジネス分析者は毎朝職場に到

着すると、自分のコンピューターにログオンします。そして、夜の間に届いたメッセージを見るために電子メール・システムを起動すると同じように、夜間のマイニング結果を見るために、データマイニング・システムを起動します。それから、データマイニング・システムが提供する情報に基づいて意思決定を始めます。

データマイニングは、短期間に好結果の得られる手法であることが明らかになったため、この分野全体の可能性に対する期待はいつそう高まっています。この分野が新しいということは、初心者の失敗リスクを少なくするための適切な資源は、まだ提供されていないことを意味しています。その資源とは、教科書、データ分析の方法、データマイニング手法の選択基準、データマイニング処理からもつとも「興味深く」適切なパターンを選択する手法、そして結果を解釈し選択するための指針などです。データマイニングを行うときに、企業が出会う基本的な問題を扱うための資源に対する要望は非常に大きく、データ量やテクノロジーだけでは十分ではありません。本書によって、著者らは一つの貴重な資源を提供しており、企業は、データマイニングの基本事項について学び、データマイニングの適用方法に関する質問の答えを得るよ

うになり、さらに先行グループの成功例や失敗例から学ぶことができます。

北アメリカIBM
グローバル・ビジネス・インテリジェンス・ソリューション
副社長

エバンジェロス・シモウディス

(Evangelos Simoudis)

まえがき

本書を手にとつて、最初の数ページに目を通してくださつていただくことに感謝します。読者が注意を向けてくださったので、本書と本書の著者たちについて少しお話しさせていただきます。

本書は、これまでの多くの本と同じように、「ギャップ」を埋めるために書かれました。そのギャップとは、データマイニングというトピックスに関して、広範囲に広がりつつある資料のギャップのことを意味します。一方では、一般的な業界誌やインターネット上で、このトピックスに関して常々報道されるようになっていますが、他方では、世界中の研究センターや大学から出される、高度な技術的および学術的な資料がおびただしく存在しています。

本書は、一般の情報よりも少し深く知りたいけれど、もつと技術的または学術的な書物を読む準備はまだできていない読者のために書かれました。本書は、この新しく興味深いトピックスについて、内容の濃い会話ができるようになりたい方であれ、あるいはデータマイニングを企業に導入するよう任されたばかりの方であれ、多くの異なるタイプの読者の興味をそそのものとなるでしょう。とりわけ、本書はデータマイニングの初心者を対象としており、複雑な数学的、統計学的な記述は用いないように心掛けられています。それゆえ、情報技術（IT）管理者ばかりでなく、情報技術に関心や基礎知識のある、多くの管理者の方々も容易に読みこなすことができます。

本書の構成

本書の構成は次のようです。

第一部 概念

第一部は、すべての読者を対象としており、データマイニングの背景にある基本概念を紹介し、現実のビジネス問題を解決するため、最新テクノロジーの主な使用法を概観します。

第一章 データマイニングの概観

本章では、データマイニングの経営面と技術面の推進要素を概観し、データマイニングを定義して、ビジネス問題を解決するデータ中心型の全体的枠組みにおける、データマイニングの位置付けを行います。

第二章 経営への適用

本章では、現在、データマイニングが適用されている、主要な経営分野について説明しています。一般的な記述を裏付ける、多数の実例を紹介し、最後にデータマイニングの適用時に潜んでいる危険に関して言及します。

第二部 基礎

第二部は、データマイニングで実際に何が行われるのか、アルゴリズムがどのように働くのか、データマイニングのベンダー・ソリューションをどのように評価するか、それを理解したい読者を対象とします。

第三章 データマイニングの処理

本章では、データマイニングのプロセスを詳しく扱っています。まず、一般的プロセスを段階に分けて紹介し、それぞれの段階の目標、だれが何をするか、どんな種類の事柄が失敗を招くか、そうした事

柄を避けるため、いくつかのヒントを提示しています。さらに、要点を示すため、この章全体にわたる数々の実例を載せています。

第四章 アルゴリズムの説明

本章では、一般的な基礎技術を持っている読者を対象とし、種々のデータマイニング・アルゴリズムを説明するため、一般的な枠組みを紹介しています。それぞれのアルゴリズムの、一般的特性、長所、および短所を詳しく解説しています。

第五章 ベンダー・ソリューションの選択

本章は、データマイニングのベンダー・ソリューションを評価する際、読者が何を調査する必要があるかを概説しています。また、データマイニング・ツールとデータマイニング・アプリケーション、およびサービスについて言及しています。

第三部 実践

第三部は、データマイニングの実施に焦点を合わせ、最初に実施例を説明し、次に読者がデータマイニングを始める方法に関して、種々の助言を与えています。

第六章 事例研究

本章では、データマイニングの現実の事例を二つ

取り上げ、詳細に説明しています。ここでは、マイニング・プロジェクトの目標、データマイニング・チームがとる方法(画面例を含む)、およびプロジェクトの結果について詳しく言及しています。

第七章 データマイニングの開始

本章では、實際上、今からデータマイニングを始めたい読者のため、評価チェックリストを示します。最初に、データマイニングの主な問題を検討し、プロジェクトの選択とコストに関して言及し、最後にデータマイニング・プロジェクトを成功させるためにいくつかの重要な要件について示します。

付録

- A IBMのデータマイニング・ソリューション
- B 特記事項
- C 参考資料
- D 用語集
- E 省略形のリスト

謝辞

本書は、多くの方々の協力があつて、はじめて仕上げることができました。このタスク全体を通して、援助や

助言や激励を与えてくださった、マイク・シャノン(Mike Channon)、ブルース・ファガティ博士(Dr. Bruce Fogarty)、およびフィリップ・ミュラー(Philippe Muller)の諸氏に對して深く感謝いたします。

さらに、初期の草稿や寄稿を校閲する点で貴重な助言を与えてくださった、以下の方々に對して感謝いたします。ラケッシュ・アグロウワル(Rakesh Agrawal)、アンドリース・アーニング(Andreas Arning)、チャック・バラード(Chuck Ballard)「付録A「IBMのデータマイニング・ソリューション」のIBMのVisual Warehouse」ジュー・ビガス博士(Dr. Joe Bigus)「第四章「アルゴリズムの説明」におけるニューラル・クラスター分割の説明」アビジット・チャタージ(Avijit Chatterjee)「付録A「IBMのデータマイニング・ソリューション」におけるIBMのParallel Visual Explorer」スザンヌ・タクス(Suzanne Dirks)「ジェラルド・ヘンケル(Gerhard Henkel)「アーマン・ト・ハースコビッチ(Armand Herscovici)「チャールズ・ヒュート博士(Dr. Charles Huot)「付録A「IBMのデータマイニング・ソリューション」におけるIBMのText Navigator」ジョージ・ジョン博士(Dr. George John)「第四章「アルゴリズムの説明」における大部分の図」ラッ

サ・リー (Russ Lee) 、クリストフ・リンゲンフェルド
(Christoph Lingenfelder) 、デビッド・マーチン (David Martin) 、
ハモウ・メスタファ (Hammou Messaïda) 、バーニス・ロゴウ
イツ (Benice Rogowitz) 「付録A」IBMのデータマイニ
ング・ソリューション」におけるIBMのDiamond」
ミシエル・ロスマン博士 (Dr. Michael Rothman) 、リカルド・
ルコ (Ricardo Rucce) 「第一章の節」リスク管理への適用」
におけるリスク管理適用例」アミット・セス (Amit Seth) 、
リチャード・シャーマン博士 (Dr. Richard Sharman) 、エバン
ジェロス・シモウデイス博士 (Dr. Evangelos Simoudis) 「序文」
ベティ・タナ (Betty Thana) 、およびアレックス・ゼクリ
ン博士 (Dr. Alex Zekulin) 「第三章」データマイニングの処
理」における原典」

「厚意により、名前の掲載および事例研究の詳細につ
いて、使用許可を頂いたHICC (Health Insurance Commission of
Australia) 、およびメロン銀行 (Melion Bank) に対して、特別
に感謝の意を表します。とくに、HICCのサイモン・ホ
ーキンス氏 (Simon Hawkins) およびメロン銀行のピータ
ー・ジョンソン氏 (Peter Johnson) のご協力を感謝いたしま
す。

最後に、制作チームのトーマス・ビルフィンガー

(Thomas Billfinger) 、マギー・カトラー (Maggie Cutler) 、ジェー
ムス・グウィン (James Gwyn : Prentice Hall社) 、バーバラ・ア
イサ (Barbara Isa) 、イブリン・ジャクソン (Evelyn Jackson) 、
ミシエル・ミーハン (Michael Meehan : Prentice Hall社) 、マリッ
サ・ステアーズ (Marissa Stears) 、マリッサ・バイベロス
(Marisa Viveros) 、およびアミー・ボーク (Amy Voge) の方々に
感謝いたします。

Peter Eisen
ピーター・キャベナ

目次

序文

まえがき

本書の構成

謝辞

第一部 概念

第一章 データマイニングの概観

バック・トゥー・ザ・フューチャー

なぜ今なのか

経営環境の変化

推進の要素 8 / 実現の要素 10

定義

改革が漸進か

第二章

経営への適用

マーケット管理への適用

カタログ電話サービスの改善 30 / ロイヤルティ・

カードによる顧客の選別 30 / 外的影響を利益へ転換

31 / 効果的な販売促進 32

リスク管理への適用

金融先物取引の予測 36 / 競争の激しい市場での価格

戦略 37

詐欺管理への適用

不適切な医療行為の検出 38 / 電話詐欺の検出 39

将来の適用分野

テキスト・マイニング 40 / ネット・マイニング 41

うまくいかないケース

離婚女性のクラス 41 / 大切な点の見落とし 42 / ダ

イエット飲料の奇妙な結果 43

どこがそれほど違うのか 17 / それほどの相違はない

20

データウェアハウスとの関係

データウェアハウス 21 / データ・マート 23 / デー

タウェアハウスからデータマイニングへ 23 / データ

マイニングからデータウェアハウスへ 23

データマイニングと

ビジネス・インテリジェンス

これからどうなるのか

15 13 5 5 3 3 xi ix ix v

41 39 37 33 28 27 25 24 21

第二部 基礎

第三章 データマイニングの処理

はじめに

プロセスの概説

プロセスの詳細

ビジネス目標の決定 51 / データの準備 53 / データマイニング 63 / 結果の分析 64 / 知識の理解 67

第五章

ベンダー・ソリューションの選択

テクノロジーの価値

データマイニング・ツール

データマイニング・ツールのタイプ 103 / データマイニング・プロセス・サポート 105 / 技術上の考慮事項 110 / 結論 113

データマイニング・アプリケーション

汎用アプリケーション 113 / 業界固有のアプリケーション 114 / 結論 114

データマイニング・サービス

コンサルティング・サービス 115 / 統合サービス 115 / コンサルタント・サービス 116 / 関連サービス 116 / 結論 117

第四章

アルゴリズムの説明

アプリケーションからアルゴリズムへ

データマイニング機能

予測モデリング 72 / データベース・セグメンテーション 74 / リンク分析 77 / 外れ値の検出 78

データマイニング手法

予測モデリング・クラス判別 79 / 予測モデリング・予測 86 / データベース・セグメンテーション・デモグラフィック・クラスタリング 87 / データベース・セグメンテーション・ニューラル・クラスタリング 90 / リンク分析・相関関係の抽出 92 / リンク分析・時系列パターン抽出 95 / リンク分析・類似時系列パターン抽出 97 / 外れ値の検出・視覚化 98 / 外れ値の検出・統計 100

第三部 実践

詐欺と悪用の防止

背景 122 / ビジネス目標の識別 122 / データの準備 123 / データマイニング 125 / 結果の分析と知識の理解 128 / 発見と利点の要約 130

ダイレクト・メールの反応率の向上

背景 131 / ビジネス目標の識別 131 / データの準備 132 / データマイニング 135 / 結果の分析と知識の解釈 138 / 発見と利点の要約 141

事例研究

122 121

131

114

113

103 102 101

79

72

69

69

51

48

47

47

第七章

データマイニングの開始

データマイニングの準備完了ですか

挑戦となる問題

社会的問題 145 / 経営上の問題 146 / 技術上の問題

147

方法の決定

ビジネス・ケース

候補アプリケーションの選択

社内実施か外注の選択

ベンダー・ソリューションの評価

技術と期間

成功要因

結論

監訳者あとがき

161 160 159 157 156 154 153 151 148 145 143 143

付録

A IBMのデータマイニング・ソリューション 11

A.1 データマイニング・ツール 11

インテリジェント・マイナー 11 / インテリジェン

ト・マイナーのユーザー・シナリオ 16

A.2 関連製品 25

Intelligent Decision Server 25 / Visual Warehouse

26 / Parallel Visual Explorer 29 / Diamond 29 /

Visualization Data Explorer 30

A.3 データマイニング・アプリケーション 30

汎用アプリケーション 30 / 業界固有のアプリケーシ

ョン 31

A.4 データマイニング・サービス 32

コンサルティング・サービス 32 / 統合サービス 33 /

教育サービス 33 / 関連サービス 33

A.5 新規出現のテクノロジ（米国） 34

テキスト・メディア・マイニング 34 / インターネッ

ト・マイニング 35 / その他 35

B 特記事項 37

C 参考資料 39

D 用語集 43

E 省略形のリスト 51

索引 3

目次

図1	データウェアハウジング・プロジェクトの平均投資利益率	9
図2	新たな顧客関係に手が届かない	11
図3	データマイニングの位置付け	13
図4	従来のデータ分析はデータマイニングではない	18
図5	データマイニングとビジネス・インテリジェンス	25
図6	データマイニングの適用領域	27
図7	類似特性を持つ顧客セグメント	31
図8	財政処置によって左右される異なる投資行動	32
図9	オンライン・データベースから抽出された特許参照情報	35
図10	母集団の体温分布の異常	42
図11	データマイニング・プロセス：ビジネス目標に始まりビジネス目標に終わる	48
図12	各データマイニング・プロセス・ステップに必要な作業量	49
図13	データマイニング・プロセス：CVAの例	53
図14	散布図：年齢と収入	57
図15	箱型図：男性と女性の収入分布	57
図16	CVAの例：サンプル・ルール出力	67
図17	データマイニングのアプリケーション、機能、および手法	70
図18	予測モデリング	73
図19	セグメンテーション	75
図20	パターンの照合	78
図21	二分岐決定木	80
図22	ニューラル・ネットワーク	83
図23	モデルの有効性：コンフュージョン・マトリックス	85
図24	線形回帰の欠点：非線形データ	88
図25	線形回帰の欠点：異常値	88
図26	相関ルール	92
図27	時系列パターン抽出：トランザクション・データベース	96
図28	時系列パターン抽出：顧客の時系列	96
図29	時系列パターン抽出：サポート値 > 40%	96
図30	5次元空間における詐欺行為の確率	99
図31	詐欺と悪用の事例研究：処理フローの全体図	124
図32	詐欺と悪用の事例研究：ニューラル・セグメンテーションへの入力	124
図33	詐欺と悪用の事例研究：確信度 50% サポート値 1%の相関関係	126
図34	詐欺と悪用の事例研究：PEIコードの除去	127
図35	詐欺と悪用の事例研究：確信度 50%のルール	127
図36	詐欺と悪用の事例研究：GPのプロファイル	129
図37	ダイレクト・メールの事例研究：融資商品のライフ・サイクル	133
図38	ダイレクト・メールの事例研究：ビジネス目標	133
図39	ダイレクト・メールの事例研究：アプローチの概略チャート	135
図40	ダイレクト・メールの事例研究：予測モデルの訓練段階	136
図41	ダイレクト・メールの事例研究：決定木の作成	137
図42	ダイレクト・メールの事例研究：リフト図	138
図43	ダイレクト・メールの事例研究：決定木による分析	140
図44	ダイレクト・メールの事例研究：決定ルールの考察	140
図45	データマイニングの開始	148
図A.1	インテリジェント・マイナーのアーキテクチャー	14
図A.2	インテリジェント・マイナー：ユーザー・シナリオの概略チャート	17
図A.3	インテリジェント・マイナー・メイン・ウィンドウ	18
図A.4	データ・ソース/ターゲット・ファイル指定	18
図A.5	欠損値のコード化	19
図A.6	「インテリジェント・マイナー・クラスタリング」ウィンドウ	20
図A.7	3つの大きなセグメント	22
図A.8	最大セグメント(33%)のズームイン	23
図A.9	Commute Distanceのズームイン	24
図A.10	最大セグメントの統計	24
図A.11	IDSカプセルの概念	25
図A.12	IDSとインテリジェント・マイナー	26