連載

データマイニング実践家達の声（1）

# データマイニング実用化：概観と展望

## Deta Mining Applications: Overview and Prospect

鷲尾　隆　大阪大学産業科学研究所
Takashi Washio　The Institute of Scientific and Industrial Research, Osaka University.
washio@ar.sanken.osaka-u.ac.jp, http://www.ar.sanken.osaka-u.ac.jp/~washio/washprjp.html

### 1. データマイニング実用は本格化へ

データマイニング技術に関する研究は1990年代半ばから本格化し，現在，世界的に一層活発化している．アルゴリズムや新分野データに関する原理や技術上の研究は，今後も大きな進展を見せることが見込まれる．さらに，データマイニングに適したデータ表現，知識表現，システムアーキテクチャ，データの前処理・後処理に関する原理や技術の研究開発など，研究領域の一層の拡大も期待される状況にある．

それと並行して，この2～3年で産業界や他学術分野でのデータマイニングの利用が本格化し，その適用領域は現在急激な拡大期にある．そしてこれら実適用分野での経験が，今後ますますデータマイニング研究にフィードバックされると予想される．データマイニング実適用の状況を知ることは，ユーザのみならず，原理や技術の研究者にとっても重要性を増していく．これから数回に亘る連載で，産業界や他学術分野の方々から，データマイニングの実用例，ないしはそれに関する批評，実用的見地からみたデータマイニングシステムアーキテクチャなどを解説いただく．本論では，その開始にあたりデータマイニング実用化現状の概観と展望を行う．

### 2. データマイニング実用化の概観

表1はデータマイニング実適用事例を大まかにまとめたものである[Washio 01]．適用事例は日を追うごとに増え，また重要機密に属する産業業務に適用され非公表のものも多くなってきている．そのため，この表自体が最新の適用現状を表しているとは言いにくい面があることを断らなければならない．

最も多くの適用事例が見られるのは金融分野である．これはマーケティング分野と各種金融業務に特化した分野に大別される．米国では，1994年頃から流通業や金融業でデータマイニングの事例が報告されているが，日本でも近年は多くの事例が報告されている．この分野では，ニューラルネットワーク，コホーネンネット，クラスタリング，分類決定木，ラフ集合，重回帰分析など，多様なデータマイニング技術が用いられている．データマイニング適用の有効性は各事例ごとに差異があるが，有用な結果が得られた事例も多い．マーケティング分野では，膨大な顧客リストから候補を見い出す必要がある．この条件下で，生命保険の潜在的解約候補顧客や効果的なダイレクトメール宛先候補顧客のマイニングでは，業務の効率の改善効果が得られている．また業務特化分野では，与信審査半無人化ルールの適用による消費者ローン無人申込機の開発や，膨大なクレジットカード使用記録からの不正利用パターン発掘において実績を上げている．

流通分野では，小売部門のマーケティングのためのデータマイニング適用が主流であり，POSデータを用いた流通全般の業務知識の導出，インストアでの販売促進用知識の導出，有望顧客の洗い出しなどが行われている．データマイニング技術としては，分類決定木，バスケット分析，重回帰分析，相関解析などが用いられている．さまざまな試みが行われているが，売れ行き予測などに対する適用事例は十分成功しているとはいえない．優良顧客の発掘や各種パターンの発掘・分析などでは効果が上がっている事例が多いが，新しいデータマイニング技術よりも全体的傾向を把握する従来の統計的手法に依拠する事例が多い．これは金融分野に比べて扱う商品や小売条件，顧客行動パターンがはるかに多様であり，顧客や購買事例を把握容易な形で類別して特徴を発掘することが難しいためであると考えられる．

製造分野におけるデータマイニングの適用は，ほかの分野と同様に進展を見せているが，多くは社内の文書やマニュアル検索，マーケティングへの適

表1　各種分野の代表的適用事例

| |
|---|
| 金融分野 |
| ・マーケティング分野<br>潜在的な住宅ローン申込み顧客の推定<br>銀行商品の適切な組合せの設計提示支援<br>生命保険の潜在的解約候補顧客の発掘<br>効果的なダイレクトメール宛先候補顧客の発掘<br>・業務特化分野<br>消費者ローン与信審査の半無人化ルールの発掘<br>リスク細分型の自動車保険の設計提示支援<br>証券顧客と営業マンとのトラブル予測<br>社債格付け推測<br>クレジットカードの不正利用パターン推定 |
| 流通・小売分野 |
| 薬局チェーン販売データからの優良顧客の発掘<br>売れ行きデータに基づく新製品販売予測<br>新製品のヒット要因分析，品物の売れ行き要因分析<br>消費者購買行動パターンの分析<br>種々の販促条件下における併売パターンの分析 |
| 製造分野 |
| HPでの顧客意見収集による次世代新製品開発<br>顧客クレーム情報による設計・製造現場品質管理<br>製造条件と製品検査結果による製造工程の改善 |
| 通信分野 |
| HP閲覧情報からの顧客プロファイリングと傾向分析<br>電話回線網管理のための負荷状況把握や障害診断<br>電話網マーケティングのための通信トラフィック分析<br>顧客の通話パターンによる通話回線不正使用検出<br>計算機アクセスログに基づく不正アクセス検出 |
| 製薬・医療分野 |
| 化学化合物分子構造と生理活性の相関解析<br>遺伝子発現形態と生理学的効果の相関解析<br>科学的根拠に基づく医療(EBM)の基礎知識獲得 |

用であり，ほかの業種と似通った目的，技術適用となっている．表1には製造業固有の適用事例を掲げた．最初のものはカスタマーリレーションマーケティングへの適用であり，主要電機メーカ，家電メーカが試みている．後2者のような品質や工程管理への適用は，広範な適用可能性を有しかつ現状も実用化が進められている．このような適用事例では，事例ベース検索やテキストマイニング，バスケット分析，分類決定木などの最新のデータマイニング技術が用いられ，効果を上げている．

通信分野では，主にインターネット網や電話網管理の分野にデータマイニング技術が用いられている．使用技術は分類決定木，バスケット分析，ベイジアンネット，ニューラルネット，テキストマイニング，サポートベクターマシン，各種統計的手法など多岐にわたる．通信分野には豊富な電子化データ蓄積があるので，データマイニングの適用可能性は高い．ネットのトラフィック診断やサーバへの異常アクセス検知，偽装チップによるなりすまし不正電話使用の検出など，膨大な通信ログから特徴的パターンを発掘する適用は成功を収めている．

製薬や医療の分野でのデータマイニングの適用は，まだ緒についたばかりであるが，早くも有望な見通しを示唆する成果が出始めており，大規模調査プロジェクトなども進行中である．化学工学や製薬の分野では，化学化合物分子構造と人体に対する生理学的影響である生理活性との相関解析へのデータマイニング適用に注目が集まっている．また，遺伝子発現形態がさまざまなタンパク質合成回路を通じて生理学的効果を表す過程は極めて複雑であり，その解析の補助や中間過程をブラックボックスとみなして遺伝子発現形態と生理学的効果の直接の相関解析を行う際に，データマイニングを使用する方法も取られている．さらに医療分野では，これまでの個別の医師の座学や経験に基づく治療内容の決定，実行ではなく，より広範な論文や調査結果，治験情報などの科学的根拠に基づき治療の実施を行うEBM（Evidence Based Medicine）が新しい潮流となってきている．その際，科学的根拠を与える知識の多くを，過去の論文や治療・治験記録データなどを調査してまとめなければならない．膨大かつ日進月歩の医療技術進歩を反映したデータから，このような知識を収集するためにデータマイニング技術が注目され，各種大規模プロジェクトにおいて適用されつつある．

## 3. データマイニングシステムの概観

上記のようなデータマイニングの実用化と並行する形で，さまざまなデータマイニングシステムが開発されてきた．これらを「機能」によって大きく分けると，総合型と専門型のシステムに分けられる．総合型は，一般的なデータベースやテキストドキュメントアーカイブからさまざまな種類のデータを取り込んだり，連携する機能を併せもち，多様なデータ前処理，データマイニング手法，後処理機能を有する汎用かつ比較的大規模なシステムとなっているものが多い．通常，データマイニングはデータを解析してみなければ，有用な知識や予測モデルなど必要な結果が得られるか否かわからない．したがって，データマイニングのために，ゼロから新しいシステムを構築することはコスト的にも時間的にも引き合わない．その点で，最初から網羅的な機能を有する総合型のデータマイニングシステムを使用すれば，さまざまな手法，処理を試行錯誤的に適用して，必要な結果が得られるか否かの検討やその性能検証などを，比較的容易に行うことができる．このため，現状実用段階で使われるデータマイニングシステムの大半が総合型となっている．これに対して専門型は，日本語テキストマイニングや地理情報など時空間的に特殊な構造を有するデータのマイニングなど，総合型で網羅しきれない特化された処理に関するシステムである．特に世界的に広範に使用されている総合型システムの中には，日本語テキスト処理や特殊用途向けの処理機能が弱いものが多い．そのようなニーズに対して特化し，高性能な処理機能を提供するものが専門型といえよう．

一方，データマイニングシステムを「用途」の観点から分類すると，スタンドアローン型，意思決定支援型，ソリューションカスタマイズ型，インテグレーション型の四つに大別されると思われる．スタンドアローン型は，データマイニングシステムとして単独で提供されているもので，その利用方法やデータマイニングスキームの構築についてはユーザサイドに任されている．ただし，データマイニングはデータ処理や解析手法，マイニングスキームの設計にスキルが要求されることがほとんどであるため，スタンドアローン型は主に研究者によって使われることが多い．意思決定支援型はデータマイニング機能よりも，ユーザとのインタラクションによって柔軟にわかりやすいデータ表示を行うシステムであり，データの大まかな傾向を直感的につかみやすくすることで，データに基づくユーザの意思決定を支援するものである．本格的なマイニング技術を用いないため詳細な解析には向かないが，マイニング技術に関する知識があまりなくても，定性的な判断が行えるメリットがある．これに対してソリューションカスタマイズ型は，上記総合型，専門型いずれでも多く提供されているシステムであり，データマイニングシステムのみならずユーザサイドの問題や解析目的に適応したデータ処理，解析手法，マイニングスキームをコンサルタントがついて設計し，ユーザと共同作業の中で目的を達成しようとするものである．産業界などにおいてデータマイニングを既存の業務に役立てる方法として，現状この形態が最も一般的であると思われる．最後のインテグレーション型は，大規模な業務基幹運用システム構築の中でデータマイニングが必要とされる場合に，基幹運用システムの1パーツとしてデータマイニング機能を組み込む用途のためのものである．これは大企業などが大掛かりな業務改革や基幹運用システムの入替えを行う際に，総合的コンピュータベンダーが必要に応じてシステムインテグレーション内で用いる．

もう一つの分類軸として，「商用ソフ

トとオープンソースフリーソフト」の観点がある．これはコンピュータOSなどにも見られる分類であるが，特にデータマイニングにおいてこの分類軸は重要である．データマイニングは，実際に適用してみないとどれだけ有効な結果が得られるかが不確定な技術である．通常の業務システムのようにあらかじめその導入メリットを見積もることが難しく，初期から多額投資を行うことはハイリスクである．その意味でシステムとしてはフリーソフトを導入し，限られた投資予算をコンサルティングに振り向けるほうが効率が良く，リスクも低減できる．機能的にも総合型のフリーソフトが登場しており，現実の多くの応用に適用可能になりつつある．ただし，コンサルティングを受ける環境面では，フリーソフトを熟知したコンサルタントがまだ少ないために，その普及には時間を要すると思われる．

## 4. 研究開発と実用化のインタラクションに向けて

これまで述べてきたように，データマイニングは本格的実用期を迎えており，かつ新たな原理や技術も次々に登場する発展期にある[Washio 03]．これは技術の発展期と実用期がオーバラップしたLSIやプログラム言語などの分野と類似しており，開発された技術シーズが新たな実用分野を拓く一方で，実用ニーズが新たな研究課題を提供するというら旋的な発展をもたらしつつある．このような状況でデータマイニング技術が健全な進歩を遂げるためには，研究開発者が常に新たな発想に基づく技術シーズを提供すると同時に，ユーザサイドからのニーズをくみ取りそれを研究に生かす姿勢が大切である．また，ユーザサイドもデータマイニング適用事例を非公開のままにせず，公開可能な限りにおいて研究開発者に事例やニーズの情報を提供すべきである．このような双方間の情報のやり取りの上にのみ，ら旋的発展が実現され，それが研究開発者，ユーザのみならず，社会全体の利益を高めていくと考えられる．現状，いくつかの学会や国際会議などにおいて，このような研究開発者とユーザ間のコミュニケーションの場が設けられているが，必ずしも十分なものとはいえない[Washio 01]．研究開発サイドおよびユーザ業務サイドの双方に不都合を生じない範囲で，実際的なシーズやニーズの情報交換が行われる場の一層の整備が課題であり，それこそがデータマイニングの技術と実用の新たな地平を拓く鍵を握ると考える．この連載がその一助となれば幸いである．

**謝　辞**

本企画を始めるにあたってお世話になった，大阪大学産業科学研究所の溝口理一郎教授に深謝いたします．

## ◇ 参 考 文 献 ◇

[Washio 01] 鷲尾 隆:ビジネスにおけるデータマイニングの現在・未来, 情報処理, Vol. 42, No.5, pp.467-471 (2001)

[Washio 03] Washio, T. and Motoda, H.: State of the Art of Graph-based Data Mining, *ACM, SIGKDD Explorations*, Vol.5, No.1, pp.59-68 (2003)

2004年4月2日　受理

## プロフィール

**鷲尾　隆**（正会員）

1983年東北大学工学部原子核工学科卒業．1988年東北大学大学院原子核工学専攻博士課程修了．工学博士．1988年から1990年にかけてマセチューセッツ工科大学原子炉研究所客員研究員．1990年（株）三菱総合研究所入社．1996年退社．現在，大阪大学産業科学研究所助教授（知能システム科学研究部門）．原子力システムの異常診断手法に関する研究，定性推論に関する研究を経て，現在は人工知能の基礎研究，特に科学的知識発見，データマイニングなどの研究に従事．著書に“Expert Systems Applications within the Nuclear Industry”, American Nuclear Society,「知能工学概論」：第2章エージェント（共著，廣田 薫 編，昭晃堂）など．計測自動制御学会，日本知能情報ファジイ学会，情報処理学会，AAAI，各会員．