5 ビジネスにおけるデータマイニングの現在・未来

鷲尾　隆（大阪大学産業科学研究所）
washio@sanken.osaka-u.ac.jp

データマイニング実用化に向けたコンテスト

データマイニング技術は1990年代前半に基礎研究が本格化したが，ここ2～3年の間に早くも産業界で実用的に使われ始めている．筆者がサーベイした範囲では，金融や流通などの分野におけるマーケティング調査への適用が主流であるが，このほかにも製造業や通信サービス業における品質管理や顧客管理などへの適用も進んでいる．このようにデータマイニング技術が急速な普及をみている背景として，データディジタル化やデータベース蓄積により対象となるデータが豊富に得られるようになったこと，および既存の統計的手法と新しいデータマイニング手法の多様な組合せが利用可能になったこと，このような組合せ利用を可能にする統合化マイニングツールが市販されるようになったことなどが挙げられる．

一方で学術研究の側面から，複数の研究者が共通データにさまざまな既存のあるいは新しいデータマイニング手法を適用し，各手法の特徴を比較評価しようとするデータマイニングコンテストが，1990年代半ば頃から盛んに行われるようになった．学術研究では新しいデータマイニングの原理や手法を提案するだけでも一定の成果と認められるが，研究者は自ら提案する手法の利点のみを強調する好都合なデータを評価に用いる傾向がある．そこで，各提案手法のメリットやデメリット，くせといった特徴を，共通データ提供者であると同時に利用者の立場でもある特定分野の専門家から，公正に比較評価してもらう場として，データマイニングコンテストが行われている．このような共通した一種のベンチマークデータによる各手法の比較評価は，機械学習分野で以前から行われている[1]．機械学習のベンチマークコンテストでは，交差検証法（Cross-validation）あるいはブートストラップ法を用い，トレーニングデータの学習に基づいてテストデータに対してどれだけ正確な推定が可能かを，正答率という明確な評価指標によって比較する．これに対してデータマイニングコンテストでは，共通データを提供した専門家にとってどれだけ有意義な知識を発見できたかを，専門家の経験的主観によって判定する．このため手法間の比較評価の客観性の確保よりも，利用者のニーズに立脚した評価を目指す場合が多い．

このようにデータマイニングコンテストは，学術研究の枠内でありながらも，利用者のニーズを中心とした手法評価を標榜しているという点で，データマイニング技術の産業界における実用化と多くの接点を持ち得ると考えられる．しかしながら現状では，産業界の実用目的と明確に連携してデータマイニングコンテストが実施されている例は少ない．そこで本稿では，産業界におけるデータマイニング技術の適用現状とその問題点を述べ，さらにそれに対してデータマイニングコンテストがいかなる寄与をなし得るかについて，可能性と課題を考察することとする．

データマイニング技術の産業界適用と問題点

適用の現状

ここでは，データマイニングの産業界適用の現状を概観する．公開された資料に基づくかぎり，最も多くの適用事例がみられるのは金融分野である．適用事例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．現在，金融業ですでにデータマイニングが適用された代表的事例を表-1に示す[2]．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，業務の効率と質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

金融分野
・マーケティング分野 潜在的な住宅ローン申し込み顧客の推定. 顧客に応じた銀行商品の適切な組合せ（クロスセールス）の設計提示支援. 生命保険の潜在的解約候補顧客の発掘，効果的なダイレクトメール宛先候補顧客の発掘. ・業務特化分野 消費者ローン与信審査の半無人化ルールの発掘. 顧客に応じたリスク細分型の自動車保険の設計提示支援，証券顧客と営業マンとのトラブル予測. 社債格付け推測，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. 投入時立ち上がり売れ行きデータに基づく新製品販売予測. 新製品のヒット要因分析，品物の売れ行き要因分析，牛乳販売量の予測. 消費者購買行動パターンの分析，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
ホームページでの顧客意見収集による次世代新製品開発（カスタマーリレーションマーケティング）. 顧客の製品クレーム情報と製造情報の突き合わせによる設計・製造現場への品質管理要求発掘. 製造現場の製造条件と製品検査結果の突き合わせによる製造工程の改善.
通信分野
ホームページ閲覧情報からの個別顧客のプロファイリングと顧客傾向分析. 電話回線網管理のための負荷状況把握や障害診断. 電話網使用需要マーケティングのための通信トラフィックデータ分析. 顧客の通話パターンによる通話回線不正使用検出. 計算機システムへのアクセスログに基づく不正アクセス検出.

表-1　各種分野におけるデータマイニングの代表的適用事例

による消費者ローン無人申し込み機の開発や，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．これまで公に報告されている代表的な適用事例を表-1に示す[3]．このように流通分野においても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，他の分野と同様に進展をみせているが，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用であり，他の業種と似通った目的，技術適用となっている．その結果，現状では製造業固有のデータマイニング適用事例はあまり多く見あたらない．表-1には製造業固有というべき適用事例を掲げた．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がいまだ試行段階の域を出ていない．後2者のような品質や工程管理への適用は，今後広範な適用可能性を有しかつ現状においても実用化が進められている事例である[4]．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．

通信分野では表-1に示すように，主にインターネットの顧客マーケティング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，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範囲は広い．特に不正使用や不正アクセスの検出など，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[5]．

使用技術と実施体制の実像

データマイニング産業界適用の実像については，2001年3月に開催された人工知能学会第14回AIシンポジウム「データマイニングはこう活かせ！」において，以上で述べた適用事例のいくつかに携わった産業界の方々を交えた調査，検討が行われた[6]．ここでは筆者のサーベイ，ならびにこのシンポジウムを通じて明らかになった内容についてまとめる．

以上に述べた適用事例を含め，産業界で用いられているデータマイニング技術は多様である．しかしながら，必ずしも1990年代に開発された最先端技術のみが用いられているわけではなく，伝統的な統計的手法も多用されている．どのような技術が用いられるかは，各事例の目的やニーズ，データの仕様のみならず，データマイニングツールの開発者やユーザによっても左右される．特に我が国では製造，通信分野の開発者やユーザは技術的な蓄積を有するため，市販ツールを利用するのみならず，種々の技術をテストしその中から最良のものを選択して，対象適用にカスタマイズ，チューニングしたツールやシステムを自ら構築することが多い．これは効果的なデータマイニングの実現には，単独の技術だけではなくデータの前処理や結果の後処理も含め，各事例に適

した先端，既存のさまざまな技術の組合せや条件設定を必要とすることが多いためである．これに対し金融や流通分野では，事例に適したツールをユーザ自らが開発することは少なく，製造や通信分野に比較すれば既成の市販ツールを用いてデータマイニングを実施することが多い．市販ツールの中には充実した機能を持つものも多く，そのため多様な技術の適用が行われている．

データマイニング技術の開発や使用時の体制も事例によって異なっている．先述のように製造や通信分野では，対象データ収集，技術開発やシステム開発，使用までをすべて自前で行う場合が多い．しかし金融や流通分野まで含めれば，コンサルタントや開発企業と組んでシステム開発，ツール使用，あるいはスキーム開発を行う場合もあるが，単独で市販ツールを購入しエンドユーザとしてそのまま利用する場合が多い．また対象データ収集についても，自前で収集したデータにマイニングを適用する場合が多いものの，目的によっては必要なデータを購入して分析を行う場合もある．

技術的および実務的問題点

以上の産業界における実像の中で，いくつかの問題点が浮かび上がってくる．これらの一部については先に言及したシンポジウムにおいて検討されている[6]．その1つは「データ収集ボトルネック」である．現場に蓄積されたデータを利用する際には，データが特定のマイニングを目的として蓄積されたものではないため，必ずしも目的達成に必要な情報を含んでいないことがしばしばある．このような目的とデータ内容の不整合性は，実際にマイニング分析を実行しないと明らかにできないことも多い．また場合によってはデータの取り直しが必要となってしまう．しかし，データマイニングのためにデータを収集し直すことは，コスト的に引き合わないことが多い．そのため，低コストな補足データ収集手段の確保や，既存データからの必要情報の推定手法などの技術が重要となるが，現状ではまだまだ未開拓である．データマイニングの研究やその応用ではマイニングの本体技術ばかりが注目されるが，データ収集法の十分な検討・改良，補足処理など，データ収集技術こそが成功の鍵を握っているといっても過言ではない．

いま1つはデータマイニングの実施体制において，往々にして市販ツールを購入してエンドユーザが使用すれば何とかなるという「市販ツール万能主義」が採られることである[6]．先述の通り，単純な場合を除けば効果的なデータマイニングには，豊富な知識と経験に基づいて各事例に適したさまざまな技術の組合せや設定条件を見つける必要がある．マイニングの目的や対象データの内容は事例ごとに千差万別である．現状の市販ツールは個別技術やそれを組み合わせる環境を提供するものであって，事例に即した適切な技術の結合による処理スキームや各種性能評価指標，チューニングパラメータ設定までは教えてくれない．データマイニングの実務への適用においては，エンドユーザが十分な知識や経験を蓄積するための時間や資本を投入するか，コンサルタントや開発会社と密接な連携体制を敷くなどの投資が必要となる．

この問題に関連してデータマイニング技術を提供する研究開発者側にも「研究分野の分断問題」が横たわっている．データマイニング技術は，人工知能やデータベース，統計など複数分野の基礎技術に根ざすものであるが，これら個別分野の研究開発者の連携が必ずしもうまくいっているとはいえず，技術がばらばらに提供されている傾向がある．また，データマイニングの実施スキーム全体を通じた各種技術の組合せにおける整合性や，各種目的やデータ内容に根ざした処理スキームの体系化などに関する研究もまだまだ手つかずの状態である．エンドユーザの立場に立ち，必要となる補助技術，各種判断指標，体系的マイニングスキームの蓄積などに関する研究が必要である．

データマイニングコンテストの可能性と課題

データマイニングコンテストの現状

データマイニングコンテストは，主にデータマイニング手法の学術研究と技術開発の見地から，複数の研究チームが，標準ベンチマークデータにさまざまな技術を適用して有用知識の発掘を競う企画である．またそれのみならず手法や分析スキームの利点，問題点などの特徴を明らかにし，さらなる理論や技術の改良，新テーマ発掘を行うものであり，世界各地で実施されている．表-2に関連する代表的企画をまとめる．最初のものはコンテストではないが，米国UC Irvine校においてデータマイニングテスト用に，多くの産業用ないしはそれに準ずるデータを，それを提供した分野専門家の連絡先とともに公開している[1]．研究者や技術者はいつでもこれらを利用して技術のテストや改良を行うことが可能である．

このような常設アーカイブ以外にも，表-2に示す多くのコンテストが国際会議などで定期的に行われている．特に米国のKDD国際会議で毎年行われるコンテストは，産業分野や科学技術分野の実データに関し，参加者が限定期間内に持てる技術力を総動員してデータ分析を行うもので，きわめて実践的内容のコンテストとして知られている[7]．またこれらのデータは上述のUCI KDDアーカイブにも収録されている．日本を含むアジアやヨーロ

データアーカイブ
米国UC Irvine校　UCI KDDアーカイブ マーケティングやインターネット関連，プロセス制御，人工衛星画像など多くの産業用ないしはそれに準ずるデータ，各々提供した分野専門家の連絡先情報．
データマイニングコンテスト
米国KDD国際会議　KDD CUP 利益最大化を目途としたダイレクトメール返送パターンの発掘やネットワークからの不正アクセスを最も効率よく検出する知識の発掘など，産業分野および科学技術分野の実データ．
アジア太平洋PAKDD国際会議　Discovery Challenge 医療疾患同定知識の発掘，化学物質毒性と分子構造の関係知識の発掘など産業分野および科学技術分野の実データ．
ヨーロッパPKDD国際会議　Discovery Challenge 金融分野における顧客提供サービス決定知識の発掘，医療疾患同定知識の発掘など産業分野および科学技術分野の実データ．
日本国内　人工知能学会知識ベース研究会 医療疾患同定知識の発掘，化学物質毒性と分子構造の関係知識の発掘，抗体たんぱく質特性に関する知識の発掘など産業分野および科学技術分野の実データ．

表-2　代表的データアーカイブおよびデータマイニングコンテスト

ッパにおいても，産業分野に関連するコンテストが行われている[8]．アジア太平洋を中心とするPAKDD国際会議および日本国内の人工知能学会知識ベース研究会では，年複数回のペースで産業および科学技術分野のベンチマークデータ公開およびそのコンテストが行われてきた．また，ヨーロッパのPKDD国際会議でも，年1回の割合で産業および科学技術分野に関するコンテストが行われている[9]．これらいずれのコンテストも発掘知識の精度のみではなく，データを提供した専門家がその実利的観点や興味深さの観点から，参加チームの発掘した知識の有用性を批評することに特徴がある．これはたとえ主観的ではあっても専門家に有用と感じられる知識を提供することが，データマイニングの主要目的であると考えるからである．

産業界に与える寄与の可能性と限界

上述したコンテストの多くが産業界の実データを対象とし，実問題に即した分析と技術上のさまざまな問題点の洗い出しを行う機会を提供している．このような実際的検証を通じて，既存技術の改良が行われるのみならず新たな技術開発の糸口が得られる．また現場専門家の理解様式に適した知識表現や知識内容発掘の必要性に迫られるため，実践的視点からの技術開発を進めることも可能となる．

さらにこのようなコンテストは研究者と産業界技術者やエンドユーザの接点を拡大する．これはデータマイニング技術の産業界への移転を促進すると同時に，産業界側のニーズや問題提起など，研究者の日常活動では得られない知見をもたらす．そしてこのような連携を通じ，新しい研究の動機付けや題材が発掘され，データマイニング技術の研究開発に広がりと厚みをもたらすことが期待される．

一方，現状のコンテスト内容には不十分な点も多々存在する．先述のコンテストの多くが，発掘知識に対する専門家の主観的有用性や興味深さを主な基準とした技術評価を行う．科学分野ではそのような基準のもとでデータマイニングが適用されることもあるが，産業界の実適用ではより具体的な目的や評価基準によるマイニングが求められることがほとんどである．目的に沿った知識が得られるか，そのとりこぼしや不要知識の混入がどの程度か，どれだけの利益に結びつくのかなど，事例内容に応じてより具体的，直接的基準のもとでのコンテスト評価も並行して行われるべきであると考える．

また，コンテストは種々の技術特性の相互比較を行う趣旨から，多くの参加チームが取り組める，すなわちさまざまな技術が適用できる一般的データや目的を設定しがちである．しかし，実際の産業界では適用可能な技術が限られる特定のデータや目的のもとでのデータマイニングが主流である．したがってコンテスト内容が産業界の事例特性とは必ずしも一致せず，産業界現場にとり直接に有用な知見が得られにくい．

さらには，産業界現場のデータや個人情報データの公開，そのようなデータに精通した専門家のノウハウの公開は難しい場合が多い．そのためにデータマイニングコンテストにおいても，データ収集ボトルネックが発生する．前述のUCI KDDアーカイブのような収集，公開努力もなされてはいるが，産業界の多様な現場状況を網羅するような事例の集積には程遠い現状である．

最後に多くのコンテストが直面しているデータマイニング遂行上の資源制約を指摘しておく．ほとんどのコンテストでは，インターネットを通じて事例データやその仕様情報の配布が行われ，数カ月の間に各参加チームがその分析結果を論文にまとめて投稿し，最後のコンテストミーティングで専門家から評価を受ける．したがって参加チームと専門家の間の緊密な連携時間は少ない．この理由は，専門家が日常業務に忙しくボランタリーなコンテスト評価や議論に十分な時間を避けないことや，参加チーム側も専門家の意見を反映した技術改良や分析のやり直しに十分な時間を確保できないことが挙げられる．

挑むべき課題

これまで述べてきた産業界におけるデータマイニング適用の問題点，およびコンテストの産業界への寄与に関する問題点は，3つの課題に集約される．1つは産業界の

実事例を幅広く整理，類型化し，各類型が有する事例の特性を効率的に網羅するコンテスト例題およびそれに関する専門家の集合を準備することである．これによって，より現場事例の目的やデータ内容の多様性に即したデータマイニング技術の評価，改良や新たな研究視点の発掘が可能となるはずである．2つ目はこのような産業界の類型事例に即して，個別のマイニング技術のみならずそれらの効果的な組合せスキームの研究開発を重視することである．このためには，コンテストの評価者として，データ分野の専門家以外にも各種処理技術に精通した複数の研究者を加えることが有効であろう．これらの研究者にはマイニングのスキームで用いられる各種技術の妥当性や互いの整合性を，技術的見地から吟味することが求められる．最後に産業界への寄与という観点から，今後のデータマイニングコンテストに望まれる実施形態について述べたい．現状のコンテストでは，データ公開からコンテストミーティングまでの期間が数カ月程度である．しかし，これでは参加チームと評価専門家が密接に議論しつつ分析を進めるには，時間的余裕が少なすぎると思われる．したがって，コンテストデータを常設アーカイブとするか，1年程度前もって公開し十分な時間的余裕を確保すべきであろう．さらに各参加チームと専門家の間の議論は他のチームにも公開し，重複議論をなくしかつ互いの議論が刺激し合うような工夫が必要である．

産業界におけるデータマイニングの展望

産業界におけるデータマイニング技術の利用は，成功事例が多数報告されるようになったとはいえ，まだまだ内容や人材が限定されており発展途上段階にある．しかし今後は実経験の積み重ねにより，データマイニング遂行に必要なさまざまな知識や経験の重要性が認識されるようになろう．そうなれば，優秀なコンサルタントや開発企業の成長，エンドユーザや開発技術者への知識や経験の浸透，および市販ツールの機能拡張や改善につれ，産業界への適用は一層拡大すると思われる．そしてこのような認識の広まりにより，市販ツール万能主義が消滅することを期待したい．適切なマイニングスキームの設計支援を行う技術開発や事例蓄積がなされて市販ツールに盛り込まれる一方，エンドユーザやコンサルタント，開発技術者が効果的なマイニングスキームを組む能力を養うことで，ツールと人材の両面から質の高いマイニングが可能となる．さらにデータマイニング研究分野の内容が深化，確立していくにつれ，データマイニング技術全体のあり方を念頭に研究開発を進める研究者や技術者も増加してくると予想される．そうなれば研究分野の分断問題も，長い目で見れば解消に向かうと考えられる．

しかしながら，データマイニングが産業界現場において真に有用な技術となることを当面妨げる可能性のある難題は，データ収集ボトルネックである．これはデータ欠如の問題であり，それを正攻法で埋めるためにはデータ収集コストの壁に直面してしまう．またより深刻なのはデータマイニング技術の研究者間でさえも，この問題の重要性が明確に認識されていないことである．しかし，補足データを収集する方法の整備や欠如情報を他のデータから推定する技術など，問題を軽減できる可能性も十分あり研究開発が待たれるところである．いずれにせよ現状では，市販ツールと既存データさえあれば誰でも簡単に有効なデータマイングができるわけではない．したがってブームに便乗するような誇大宣伝は謹むべきであり，現状技術の可能性と限界に関する正しい認識を持つことが必要である．

一方，学術研究と産業界現場の接点については，今後多様化が進むことが予想される．データマイニング研究は，ある目的の実現に向けて多様な技術を組み合わせる総合工学的な研究分野である．したがって学術研究と産業界実践の距離が近い分野であるといえる．現状における両者の具体的接点はデータマイニングコンテストであるが，主に研究者や産業界技術者間の接点であり，また技術情報発信の場に限られている．しかし一方で産業界のエンドユーザの声を研究に反映させる機会の提供も必要である．先に言及したAIシンポジウムは，そのような場の1つとして今後の展開が期待される[6]．産業界の実践事例の報告およびそこにみられるニーズや課題を発信し，エンドユーザと研究者，技術者がいっしょになって検討する場の重要性は増加すると思われる．

参考文献

1) Bay, S. D.: UCI KDDアーカイブ: データマイニング研究と実験のための大規模データ集合のアーカイブ，情報処理，Vol.42, No.5, pp.462-466 (May 2001)．
2) 小野　潔: 金融業におけるデータマイニングの応用，第18回日本SASユーザー会研究発表論文集，pp.159-171 (1999)．
3) (株) 日経リサーチ: POSデータに対するデータマイニング事例集 (June 2000)．
4) 堀　聡他: 電気製品の市場品質監視システムーデータマイニング技術の応用ー，人工知能学会誌，Vol.15, No.5, pp.813-820 (2000)．
5) Hashimoto, K. et al.: Probabilistic Modeling of Alarm Observation Delay in Network Diagnosis, Proc. of PRICAI 2000, pp.734-744 (2000)．
6) 人工知能学会第14回AIシンポジウム資料: データマイニングはこう活かせ！，SIG-J-A004 (http://wwwsoc.nacsis.ac.jp/jsai/AI-sympo.html) (2001)．
7) Kohavi, R. et al.: KDDカップ2000主催者報告書: 本質をつかむ，情報処理，Vol. 42, No. 5, pp.445-453 (May 2001)．
8) 鈴木英之進，津本周作: 日本・アジアにおけるデータマイニングコンテスト，情報処理，Vol.42, No.5, pp.457-461 (May 2001)．
9) Berka, P.: ヨーロッパにおけるデータマイニングコンテスト，情報処理，Vol.42, No.5, pp.454-456 (May 2001)．

(平成13年3月31日受付)