海外におけるデータマイニング事例

山端　博

1. はじめに

1993年，データマイニングの手法が，米国IBMの提唱する革新的なビジネス課題解決の方法として登場して以来，多くの年数が経過した[1]．その間，海外でのデータマイニングの活用技術は，多くの事例を生み出し，これから益々発展していこうとしているが，特に実務への適用で先行していた金融・流通などの業界では，データマイニングの応用において一日の長があり，分析応用の方法論においては既に習熟期に入っているといえよう．

それらのデータマイニング事例において，細かな分析上の差異を別にすれば，これまでに確立した共通のアプローチや方法論のようなものはあるのだろうか？実はデータマイニングのアプローチは適用する分野や業界，また対象とするデータに応じて千差万別であり，未だに専門家の経験則の中に眠る暗黙知のケースが多いのが実状である．これには，いくつかの事情があるが，一つにはデータマイニングの適用領域が多くの場合企業の戦略的な部分に活用されてきており，直接機密にかかわるケースが少なくないためである．しかし，そのような中にあっても，基本的な作法を押さえておくことは，新たにデータマイニングに取り組もうとする諸氏にとっては，大いに役立つ部分があることもまた事実である[2]．また，ハイブリッド・データマイニングの体系化など，一般的な方法論や純粋に先進的な技術論の適用において多くの有益な情報が公開されてきており，洞察力のあるアナリストであれば，容易に参考情報として取り入れ，応用に資することができよう[3]．今回は，二つの海外事例を通し，データマイニングを実務に活かす上で，多少なりともお役に立つ情報が提供できれば幸いである．

やまはた　ひろし
日本アイ・ビー・エム㈱ ビジネス・イノベーション・サービス事業部
〒103-8510 東京都中央区日本橋箱崎町19-21

2. 保険会社の事例

上記で述べたように，習熟期に入ったデータマイニング・ビジネスはアルゴリズム研究の段階を過ぎ，実務でのパフォーマンスを競う段階に入ってきている．ここでパフォーマンスという時，一つは処理能力の観点であり，データマイニング定義のキーワードでもある大容量データが高速に処理できることを意味している．例えば，米国最大手の総合保険会社ステートファームの顧客分析プロジェクトは，当時で30 TB（テラ・バイト）を超すデータウェアハウスを活用した，典型的なデータマイニングの事例といえる[4]．しかし今回の事例紹介では，規模の観点よりも，むしろアプローチそのものの特徴に焦点を当てたい．

2.1 米国保険業界におけるデータモデルの特徴

分析のパフォーマンスという時，一方で成果を意味することになるが，米国保険業界では一般的に，どのような分析がなされているであろうか．この時，その業界や各企業ごとの特徴が現れるが，米国保険業界でも事情は同じである．

一つは，データモデルの問題がある．今回は，実際の保険会社で使用されたモデルを直接ご紹介することができないため，米国IBMにおいてプロトタイプとして提供され，比較的共通性が高いと思われるデータモデルを参考にご紹介する（図1-1）．

これらのデータ項目は，新規契約顧客獲得のためのデータモデルであるが，ゴルフクラブ会員権の有無や，クレジットカードの保有状況については各カテゴリー別に把握されており，かなりの個人情報が入手されていることが分かる．しかも，これはプロトタイプであるから，本来使用されているもののイメージを与えるに過ぎず，実際は更に詳細なレベルの個人情報が収集されていると考えられる[1]．このように海外事例を見た場合，国によりデータの入手可能性についての事情が異なるため，必然的に適用可能なアプリケーションについては限界があり，違いがあることに注意

1 アカウント KEY
2 顧客:団体ID
3 顧客:年齢
4 顧客:信用ランクコード
5 顧客:現在年収
6 顧客:雇用状況コード
7 自動車保険契約顧客
8 生命保険契約顧客
9 顧客:婚姻状況
10 顧客:分析用顧客No.
11 顧客:扶養成年数
12 顧客:扶養子供数
13 顧客:過去引受謝絶
14 顧客:性別
15 顧客:郵便番号
16 顧客:州
17 顧客:貯蓄状況
18 契約:年間保険料
19 自動車保険:ステータスコード
20 自動車保険:過去引受謝絶
21 自動車:合計クレーム金額
22 郵便番号
23 世帯主
24 銀行クレジットカード
25 百貨店カード
26 小売業者カード
27 プレミアムカード
28 ライフスタイルカード
29 生活装備のハイテク度
30 消費性向コード
31 持家状況
32 持家相場価額
33 ゴルフクラブ会員
34 家屋火災保険
35 世帯内での所得者数
36 契約:証券発行日
37 契約:付加条項数

図 1-1　新規顧客獲得のためのデータモデル

が必要である．以下は，米国保険会社での事例である．

2.2　ビジネス形態によるモデル適用方法の違い

米国においても保険業では，代理店やブローカーが存在し，自社の保険商品を消費者に説明し販売しており，クロス・セリングなども行っている．一方，コール・センターなど顧客維持を主たる目的とした自社の直営チャネルでも，クロス・セリングなど既存の契約顧客に対する販売を行っている．このように性質の異なる複数のチャネルを持つ場合，その分析結果の解釈や施策への展開方法などが同じでよいかということが，ここでの関心事となる．データマイニングによる分析結果を施策に適用する時，ともするとチャネルの扱いは同じであり，それぞれにモデルを区別する必要はないと考えがちである．しかし，代理店やブローカー・チャネルの場合，通常，彼らは保険のプロフェッショナルであり，対顧客のアプローチについては，自らの業務体験に基づく経験則のようなものを持っている．そのため，彼らを通して施策を実施しようとする場合，それらの経験則を無視した顧客プロファイルによる対象者リストを与えた時に，施策担当者に受け入れられるかどうかという点で課題が残る．したがって，この場合の代理店やブローカーに渡される顧客リストは，

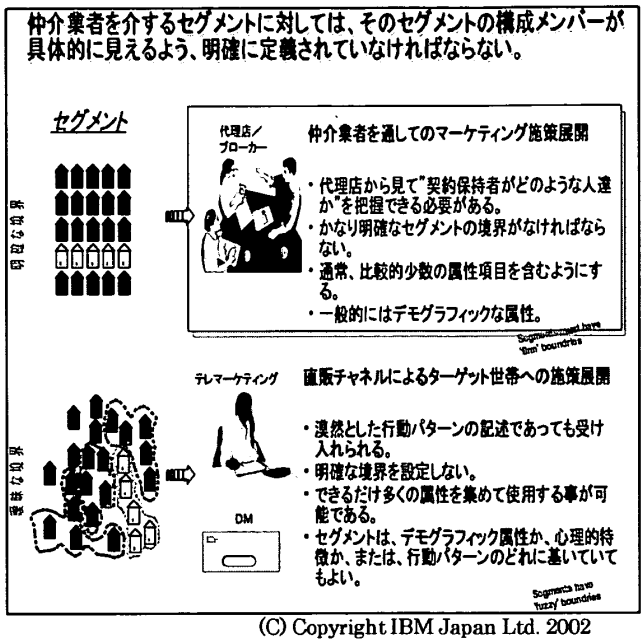

図 1-2　施策チャネルによるセグメンテーションの違い

施策担当者に理解されるよう明確な顧客プロファイルを持つ必要がある（図 1-2）．

一方，テレマーケティングのチャネルでは，コールセンター・オペレーターにより，時間当たりの生産性を重視した施策展開が行われるため，顧客のプロファイリングのような特性値による対象者の意味づけは不要である．むしろ，顧客属性を含む行動パターンや心理特性など，より科学的・客観的に精度を追求し，導き出された結果を施策対象セグメントとすることで，クロス・セリングの成果を上げることができる．

以上，米国保険業界での代表的なデータマイニングの事例をご紹介したが，分析に使用するデータ，また，分析の目的や方法など，その国のビジネス慣習が大きく影響していることがお分かり頂けたと思う．次に，米国の公共的な領域で活用されているデータマイニングの実例を見てみよう．

3.　FAMS とその応用事例

データマイニング発展の経緯が，歴史的に金融・流

[1] 筆者がたまたま見かけたあるデータモデルでは，このほかに Co-Brand カードの種別や喫煙の有無，更には前科まで入っていたものもある．これは，米国ではすべての国民が社会保険番号（Social Security Number）によって管理されており，個人を完全な形で特定できるため，外部データ業者などにより個人情報が整備され外販されているためである．これらは，国勢調査データや，企業が倒産したときに流出する顧客リストなどを含め，一般的には合法的に入手できる[5]．いずれにしても，これらの個人データは，マーケティング担当者，特に新規顧客獲得を行う施策担当者にとって有用なデータではあるが，これを個人の側から見ると知らないうちに自分のデータが売買され，分析され，自分への商品・サービスの販促手段に使われていることになり，プライバシー上の問題として議論の対象となる部分でもある．この点では EU からの圧力，および国内での批判の高まりを受け，米国でも現在アメリカ議会で法案を検討中のようである．

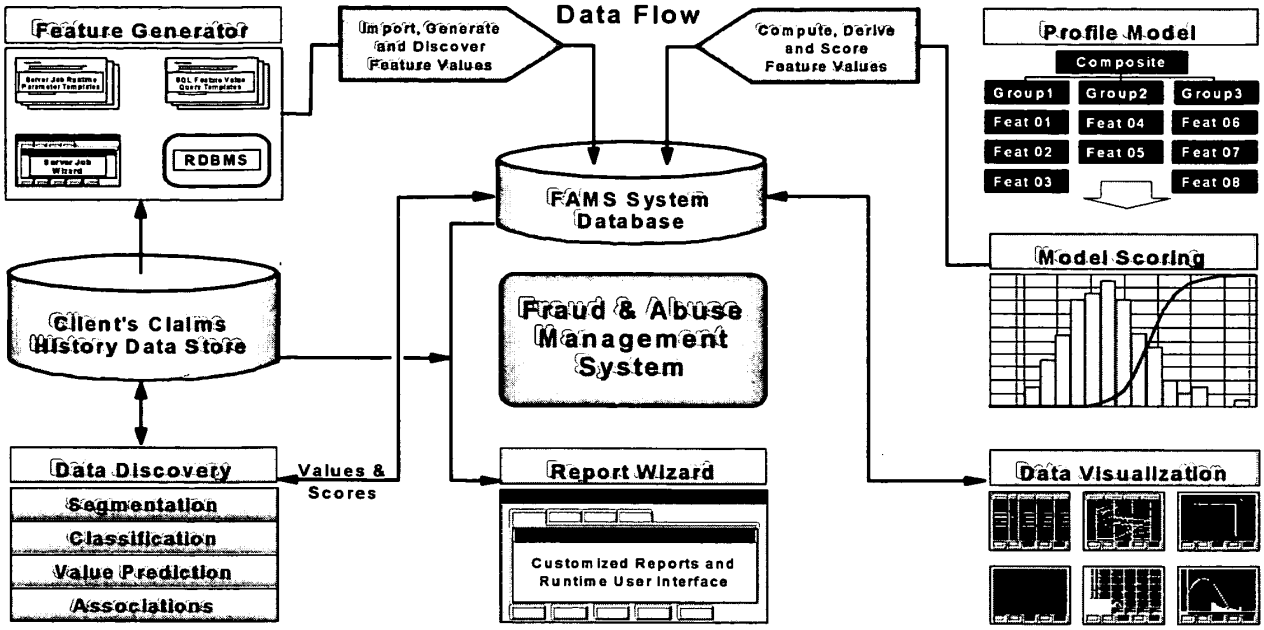

図 2-1　FAMS システムのコンセプトとデータフロー

通などの大量データの分析から始まったことにより，その適用領域が収益を主たる目的とする民間ビジネスのエリアに限られるという先入観が一般的である．実際，よく知られた代表的な適用領域を見てみると，金融業における貸付審査業務アプリケーションや，流通販売業における優良顧客のターゲティングなどが，それぞれの業界においてよく知られている．これは，データマイニングが，ROI（投資対効果）を実現する手段として発展してきたことからも，ごく自然に理解されることである．しかしデータマイニングを，科学的にデータを取り扱う意思決定のための方法論と位置づけるとその歴史は古く，データマイニングという言葉が現れる以前から米国では政府系機関や公共機関が，予算配分の公正さを期すための手段として統計的な手法を用いた政策決定を行っていた．このような背景から，データマイニングが政府系機関や公共機関においても，素早く取り入れられ活用されるのも時間の問題であった．ここではいくつかの事例の中から，特に公共性の高い FAMS（Fraud and Abuse Management System）を使用した事例をご紹介する．

3.1　FAMS の成り立ちと代表的事例

FAMS は，不正パターン発見のために開発されたソリューションである．全米で，少なくとも 20 以上の医療保険機関が利用し，医療保険の不正請求を発見するための手段として活用されている．

基本的な考え方は，不正な請求は通常のものと比べて，どこかに必ず不自然なパターンが潜んでいるという統計的な理論に基づいている．以下にその手順を示そう（図 2-1）．

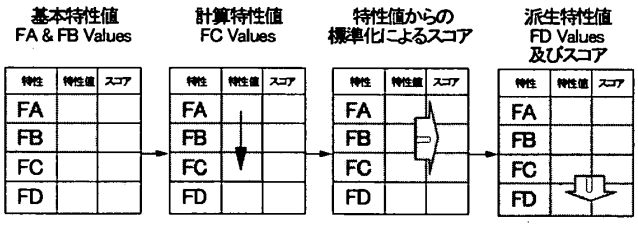

図 2-2　各特性値の生成過程

FAMS システムは，医療請求データを Client's Claims History データストアと呼ばれる履歴データベースに格納する．これは後続プロセスで，医療請求データに潜む特殊パターンを発見するためのパラメータを生成するためである．このパラメータ群の中に基本特性値（Base Feature）がある．基本特性値には，データストアより得られる患者の被請求額などの数値データのほかに，年齢性別，居住地域などのデモグラフィックな属性も含まれる．

さて，データマイニングにおいてよくいわれることの一つに，「生データをそのまま分析するな」という常套句がある．これは，ソースデータを加工・変換することで，より説明力の大きい新たなパラメータ（変数）を生成することができるからである．FAMS もこの方針に則り，基本特性値からいくつかの計算過程を経て新たな特性値を算出している．これを模式的に表すと図 2-2 のようになる（図 2-2）．

上記において計算特性値（Computed Feature）は，基本特性値の加重平均により得られる．また，すべての特性値は1から1,000の範囲で標準化されるが，これを「スコア」と呼ぶ．更に，以上の特性値とスコアのすべてから派生特性値（Derived Feature）と呼ばれる新たな特性値を生成し，同様にスコア化も行う．これにより，不自然なパターンを発見するための最終的なパラメータ群が生成されたことになる．FAMSは，これら一連のパラメータ群を分析対象（プロバイダー）のプロファイリングを行うために使用する．

プロファイリングのコンセプトは簡単である．すべてのプロバイダーは，Peerグループと呼ばれる類似のセグメントへと分けられており，すべての特性値は，このPeerグループを最も特徴的に表現するものとして対応づけられている．このPeerグループは，具体的には，ある地域を担当するある専門医（例えば小児科など）のように，明確な基準によっても設定されるが，クラスタリングやRBFなどのデータマイニングの手法により導くこともできる．このPeerグループは，本来，同質であるべき類似な集団として導かれるため，この中で大きく乖離（Deviate）するものは，

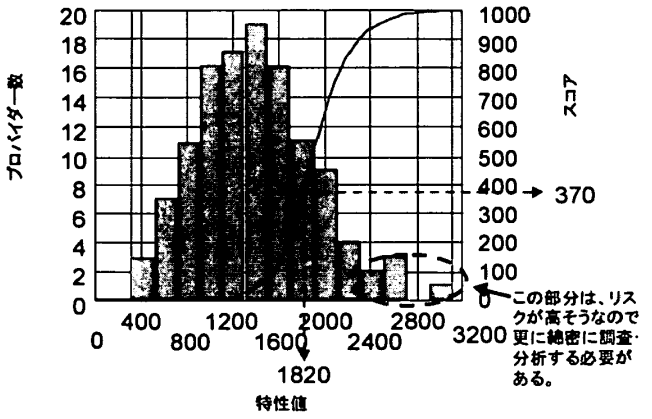

図2-3　特性値とスコアの関係

何らかの原因による不自然なパターンとして認識され，不正請求の調査対象候補としてリストアップすることが可能となる（図2-3）．

このための，乖離とリスク度の関係を定義付けるスコアの分布形が以下のように定義されている（図2-4）．

1. Growthパターン…特性値が増加する程，不自然さの度合いが増す．
2. Declineパターン…特性値が減少する程，不自然さの度合いが増す．
3. Bellパターン…特性値の両極端で，不自然さが最小となり，中央値で最大となる．
4. U-Bellパターン…特性値の両極端で，不自然さが最大となり，中央値で最小となる．

FAMSでは，これらの不自然な対象（外れ値）を，それぞれの目的に応じて見やすい形で提供されたビジュアライゼーション機能によって容易に発見することが可能である．また，各外れ値の詳細情報もレポーティング機能により具体的に確認することができる．

3.2　Tax Audit and Compliance System

FAMSの徴税システムへの応用として，TACS（Tax Audit and Compliance System）システムがある．TACSシステムの目的は，ニューヨーク州の徴税監査プログラムにおいて，その生産性を最大化することであった．まず，当システムの設計目標の第一は，高度に訓練された徴税監査人が，特に大きな追徴金を期待できる納税者にフォーカスできるようにすること．更に，当システムは，納税調査，徴税訴訟，支払完了までの全プロセスをサポートする情報を提供できることが必要な条件であった．そして最終的には，将来にわたる税収の損失を未然に防ぐための十分な情報を州税務当局に提供できることである．TACSシステムのコンセプトは，ニューヨーク州税務当局に提出され

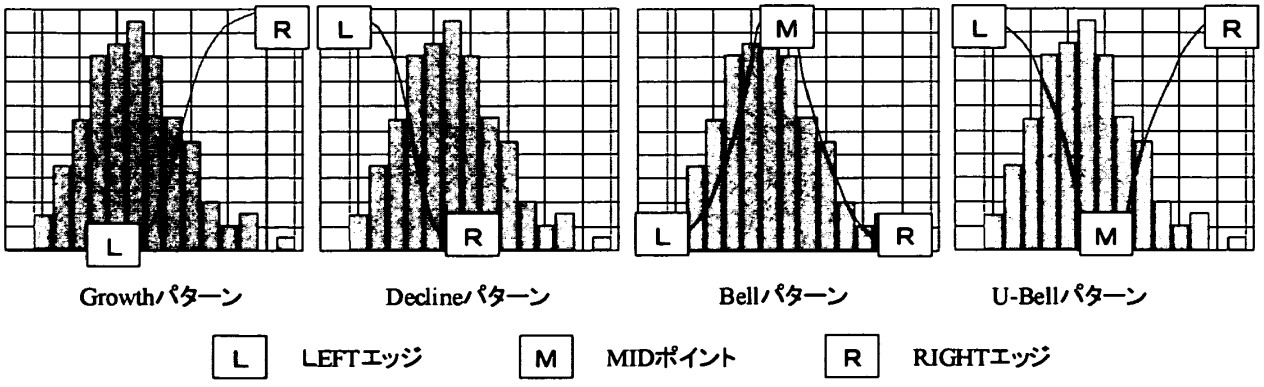

図2-4　特性値に対するスコアの分布形

た「Federal Schedule C」と呼ばれる納税申告書に基づき，不自然な偏りのある納税パターンを例外的な事象として発見・報告することであるが，実装的には州税務当局のオペレーショナル・データストアにアクセスするアプリケーションとして，包括的なデータウェアハウスアーキテクチャの中に組み込まれている．

TACSシステムでは，FAMSシステムの基本的プロセスに従い，まず，納税者をPIA（業種）コードやその他，共通の属性を保有するセグメントに分け，Peerグループとして定義する．更に，特性値抽出機能が，ニューヨーク州と連邦税務局のデータにアクセスし，各Peerグループごとに「総収入に対する純利益率」のような基本特性値を計算し，データベースにロードする．更にFAMSと同様，計算特性，および，派生特性が計算され，各々のスコア値も算出される．TACSシステムでは，これらすべての特性値とスコアをビヘイビア（Behavior）と呼び，最終的に脱税容疑の高い候補者にフォーカスし抽出するための「脱税容疑インデックス」として使用されることとなる．

さらに，TACSシステムは，現在の個人事業主の徴税業務に加え，売上税・法人税に対する徴税監査プログラムへも機能拡張しているところである．また，TACSシステムのビジュアライゼーション機能とレポート機能は，州税務当局のプロファイル・データはもちろん，連邦税務当局の徴税データにも素早くアクセスすることができるため，監査人が高リスクな納税者に的を絞り，徴税監査の生産性を著しく高めることに寄与している．

4. おわりに

以上，今回の海外事例では，二つの対照的なケースを挙げた．一つは，保険業界では世界最大規模のデータウェアハウスをデータマイニングし，収益性の向上を追求する事例であり，もう一方は，対照的に公共性が高く，不正発見の手段を提供するツールとして紹介した．データマイニングが民間企業の利益を追求するための手段としては，既にかなり定着してきたことを考えた場合，これからはデータマイニングの次なる段階として，公共性・社会性のある分野への適用が期待されるのではないかと思われる．今後益々，当分野が世の中のお役に立てるような方向へと発展することを期待して止まない．

参考文献

[1] 山端博，“ビジネスインテリジェンスとCRM，—データマイニング・ビジネスの実際—”，日本オペレーションズ・リサーチ学会，1998年12月号，Vol. 43 no. 12.

[2] 大内雅晴，“データマイニングを企業で成功させる方法”，日本オペレーションズ・リサーチ学会，2002年9月号，Vol. 47 no. 9（当月号）．

[3] 小野潔，“金融業におけるデータマイニングの応用，—保険解約の防止分析—”，日本オペレーションズ・リサーチ学会，2000年5月号，Vol. 45 no. 5.

[4] Business Intelligence Solution，IBM顧客事例紹介ブローシャ，G 588-1742-00.

[5] 菅恭二監修・三村聡・本田伸孝，金融マーケティング戦略，—銀行経営を変えるCRM—，㈱金融財政事情研究会，1999年11月．