顧客分析とデータマイニングの動向

Database Marketing and Data Mining

松 田 芳 雄

要 約 CRM（Customer Relationship Management）などの新しいマーケティングの考え方の出現により，一人ひとりの顧客の特性を識別する必要が生じ，データマイニングが注目されている．

データマイニングは，大量データから規則性や関連性などを自動的に抽出する手法である．本稿では，マーケティングにおける顧客分析の要件を整理し，この分野におけるデータマイニング・ソフトウェアの必要な機能を考える．

今後インターネットやコールセンターの普及により，リアルタイムでデータマイニングを行うオンラインマイニングや文字データを分析する文書マイニングも必要になってきている．これらの新しい技術についても紹介する．

Abstract Data Mining has lately attracted considerable attention because of the new marketing method like CRM（Customer Relationship Management）. Data Mining is an analysis method of large quantities of data in order to discover meaningful patterns and rules. In this paper, we discuss the necessary condition of data mining software from the point of view of marketing analysis. And we introduce online mining and text mining, which are new technology of data mining.

1. はじめに

最近のマーケティングは，一つの商品をできるだけ多くの顧客に売るという考え方から，一人の顧客にできるだけ多く買ってもらうという考え方に変わってきている．CRM（Customer Relationship Management）に代表されるように，顧客との良好な関係を保ち続け，顧客一人ひとりの要求に応えるような仕組み作りが行われている．そのためには，個々の顧客の違いを識別する必要があり，データマイニングの手法が注目されている．

2. データベースマーケティングとデータマイニング

2.1 マーケティング方針の転換とデータマイニング

日本では通信販売をはじめとするダイレクトマーケティングの業界は，1980 年代の後半に基幹の業務システムの整備が終了し，個人顧客の購買データを収集できるようになった．これらのデータを売上げや収益の向上に活用することが検討され始めた．通信販売会社の多くは，自社に蓄積された顧客リストから毎回 30〜40% の顧客を抽出してカタログやダイレクトメール（DM）を発送している．そこでは当然効率が重要視される．少しでも購買見込みのある顧客を選択するために，顧客一人ひとりの購買可能性を計量的に評価する方法が検討され始めた．その頃，アメリカからデータベース上に蓄積されたデータをもとにマーケティング戦略の立案や販促活動などを行うデータベースマーケティングという概念が紹介された[1]．しかし，日本の企業でマー

ケティングに利用できるデータを蓄積している企業は少なく，採用は一部の企業や業態にとどまった．

その後，景気後退の影響などで新規の顧客の獲得が難しくなり，そのための費用も増大しマーケティング方針の転換が図られるようになる．すなわち，ワンツーワンマーケティングの導入である．新規顧客を獲得するよりも，既存顧客からの売上げを拡大する「顧客獲得」から「顧客維持」へと方針が転換された．これは，従来の一つの商品をできるだけ多く売る「市場シェア」重視の考え方から，一人の顧客の消費をできるだけ多く自社に取り込もうとする「顧客シェア」重視の考え方への変更である．マーケティングの方法もテレビなどの不特定多数に対する広告を用いた「マス対応」から，個々の顧客の要求を把握し満足度を向上させる「個別対応」へと変わってきている[2]．

以上のようなマーケティング上の要求を満たすためには，一人ひとりの顧客の特性を識別することが必要である．現在，ポイントカードやマイレージカードを発行して顧客個人のデータを収集している企業は多い．こうして集めた顧客データを分析することにより売上げの拡大を図ろうとしている．そこで，顧客データから個々の顧客の特性を明らかにする顧客分析が重要になっており，そのための手法としてデータマイニングが注目されている．

2.2 データマイニングとは

データマイニングは，大量データの中から規則性や関連性など意味あるパターンを自動的に抽出する手法である[5][6]．この規則性や関連性はルールと呼ばれる．従来，データ解析は多変量解析などの統計的手法が主流であったが，1990 年代の中頃，人工知能の機械学習の分野でニューラルネットやデシジョンツリーによるルールの自動生成の研究がデータマイニングの発端である．

ここで，ルールとは以下のようなものである．〔ルール 1〕「初回に新聞チラシで家具を買う人はリピートがない」．〔ルール 2〕「男で独身で勤続 1 年未満で電話連絡がつかない人は延滞率が高い」．いかにもありそうで当たり前のような例ではあるが，これらのルールをデータから導き出し信頼性を証明するのは結構難しい仕事である．これら二つのルールは何れも 10 年程前に実際に分析した実例である．当時は汎用コンピュータの時代で，データマイニング手法の原形のようなものは存在していたが，大規模なデータには適用不可能であった．プログラムを開発したり，集計ソフトウェアを組み合わせたりして分析を行い，いくつかのルールを抽出した．実際，〔ルール 1〕には 6 人月位の工数，〔ルール 2〕には 2 ヶ月間位の期間を費やした．分析の方法は仮説検定である．いくつかの仮説を立て，データを集計することによりそれらを一つひとつ検証していったが，コンピュータが現在のように速くないこともあり多大の時間を要した．

データマイニングは仮説検定の手法ではなく仮説の発見手法でもある．通常，仮説は人間が設定するものであるが，仮説を思いつかなければこれ以上作業をすることができない．これらの作業でも度々そういう状況に陥り，別の手法でヒントをつかむようなことを行った．データマイニングは図 1 のような形で仮説（ルール）自身を表示するので，仮説を持っていなくても分析を進めることができる．図 1 はデータマイニ

ングの代表的な手法の一つであるデシジョンツリーの例である．顧客の「年齢」「勤続年数」などの属性と延滞率のデータを用意しておき，延滞率の差が最も大きくなる属性から順に顧客を層別したものを 2 分木で表現したものである．この例では，「持ち家以外に住み，年齢が 50 歳以上で，他社からの借入が 6 件以上あり，勤続 10 年未満の人が最も延滞率が高い」というルール（仮説）を提示している．

統計解析では，少量のサンプルから全体の母集団の特性を推定し，その信頼性は有意水準などの確率で与えられる．一方，データマイニングは推定の信頼性などの検定は一般に行うことは少ない．そのかわりに全データに近い大量データで実行して信頼性の保障をしようとする．仮説検定と仮説発見，少量データからの推定と大量データによる信頼性の保障が統計的データ解析とデータマイニングの違いである．最近では，伝統的な統計的データ解析手法と新しい手法とを融合させて適用するのが一般的である．

現在では，〔ルール 1〕と〔ルール 2〕のルールは，データが用意されていれば，統計解析や他の手法の専門的な知識が無くとも半日か 1 日位の作業で抽出が可能である．デシジョンツリーの手法は，データマイニング以前から同様のものは存在していたが，使われる場面はそれほど多くはなかった．少量サンプルでは信頼度に問題があるし，大量サンプルではコンピュータの能力から実行不可能であった．適用事例もほとんど無かったが，大量サンプルで実行してみると仮説発見という使い道も分ってきた．現在，デシジョンツリーはデータマイニングの代名詞のようになっており，最も使われている手法である．データマイニングは，コンピュータの高性能化とデータの分析技術の向上による新しいデータ活用技術である．

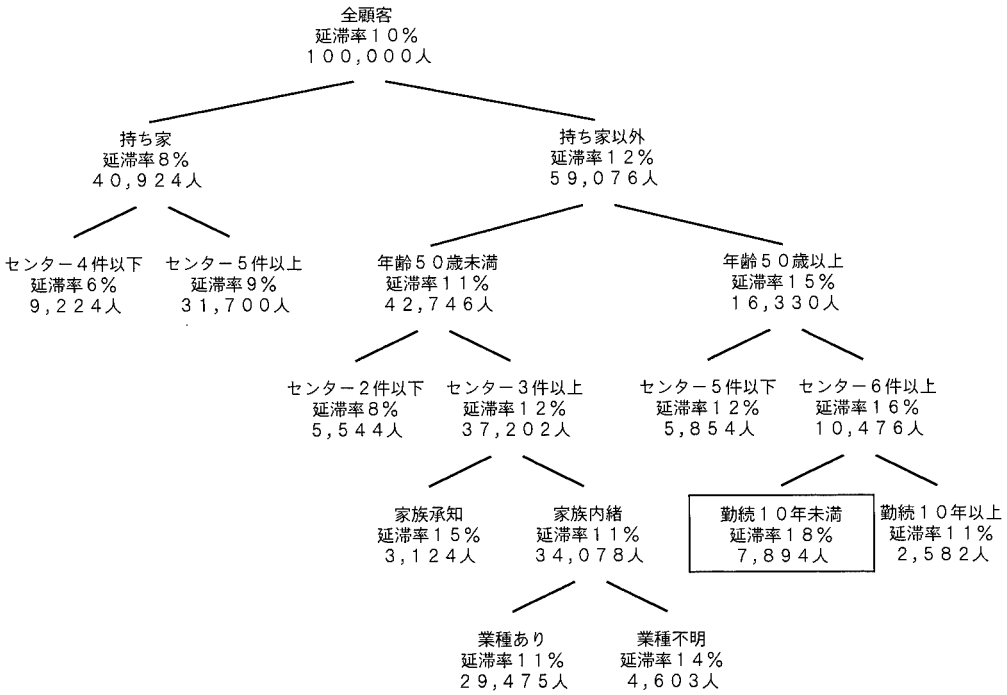


図 1　データマイニング（デシジョンツリー）の例

3. データマイニングの手法

データマイニングは唯一つの手法ではなく複数の分析手法の総称である．ここではそれらを以下の四つの手法に体系化する．

1）「予測」 ある値（連続量）の予測や Yes/No（二値）の判別を行う手法である．
2）「分類」 サンプルをいくつかのグループに分ける手法である．
3）「判別」 サンプルを分割して有益なグループを探す手法である．
デシジョンツリー（決定木）はこの手法である．
4）「相関」 二つ以上の事象の関連性を探す手法である．
バスケット分析やシーケンス分析とも呼ばれる．

3.1 「予測」

ある値（連続量）の予測や Yes/No（二値）の判別を行う手法である．連続量を予測する手法としては，回帰分析や数量化理論Ⅰ類が最も一般的である[14]．サンプルが二つのカテゴリのうちどちらのカテゴリに属するかを判別する手法には，判別関数やロジスティック回帰分析がある[12][13]．ニューラルネット（バックプロパゲーション法）での判別も行われる[15]．

3.2 「分類」

サンプルをいくつかのグループに分ける手法である．グループ分けの手法の代表的なものはクラスター分析である[16]．クラスター分析には，クラスターの数を段階的に増加または減少させる階層的な方法と，クラスターの数をあらかじめ固定しておく非階層的な方法がある．データマイニングでは，大量サンプルを分類するために，非階層的手法である k-means 法などがよく使われる．ニューラルネットではコホーネンによる方法などがある[18]．サポートベクターマシンという新しい手法も考案されている[20]．

3.3 「判別」

サンプルを分割してデシジョンツリーを作成する方法である．外的基準（サンプル分割の評価基準となる変数）や説明変数（分割を行う変数），評価基準，分割が 2 分割か多分割かの分割数，分割を停止する条件などにより多種の手法がある．表 1 に代表的な手法を示す．

表 1 デシジョンツリーの作成手法

	外的基準	評価基準	説明変数	分割数	停止条件
AID	量的	平方和	質的	2分木	停止条件（最小サンプル数や最大分割数）に達するまで行う。
CHAID	質的	χ^2値	質的	多分木	
CART	質的 量的	平方和 情報量	質的･量的	2分木	分割できなくなるまで行い、最後に枝刈りする。
ID3	質的	エントロピー （利得基準）	質的	多分木	
C4.5	質的	エントロピー （利得比基準）	質的･量的	多分木	

AID (Automatic Interaction Detector)
CHAID (Chi-squared Automatic Interaction Detector)
CART(Classification and Regression Trees)
ID3 (Interactive Dichotomiser 3)
C4.5

3.4 「相関」

二つ以上の事象の関連性を探す手法である[7]．主に顧客の併買商品などを分析する方法である．たとえば，シャンプーとリンスを考える．シャンプーを買った顧客のうち何パーセントの顧客がリンスを買うかを分析し，商品間の相関を発見する．スーパーの買い物カゴ（バスケット）の中に何と何が入っているかということからバスケット分析と呼ばれる．時系列的に時点をずらした場合をシーケンス分析と呼ぶ．昨年ハワイに行った人は今年はどこに旅行するかというような分析を行う．図2にスーパーでの併買をバスケット分析で行った例を示す．

● 牛肉（国産）との併買ベスト5

1	16.4	国産牛	佃煮、こんにゃく
2	15.6	国産牛	野菜類、こんにゃく
3	14.9	国産牛	牛乳、中華の素
4	14.8	国産牛	こんにゃく、（日曜日）
5	14.1	国産牛	豆腐、こんにゃく

● 鶏肉との併買ベスト5

1	30.9	鶏肉	唐揚げ粉
2	18.4	鶏肉	野菜類、こんにゃく
3	18.4	鶏肉	ベーコン、豚肉
4	17.8	鶏肉	佃煮、納豆、豚肉
5	17.8	鶏肉	牛乳、シチューの素

● パン粉との併買ベスト5

1	4.6	パン粉	サラダ油
2	4.0	パン粉	貝類
3	3.8	パン粉	豚肉、漬物
4	3.7	パン粉	豚肉、調味料
5	3.4	パン粉	豚肉、牛乳

● 牛肉（輸入）との併買ベスト5

1	9.0	輸入牛	焼肉のたれ
2	3.4	輸入牛	野菜類
3	3.1	輸入牛	カレールー
4	3.0	輸入牛	スープの素
5	2.8	輸入牛	香辛料

● 鮮魚との併買ベスト5

1	26.9	鮮魚	刺身
2	24.7	鮮魚	納豆、貝類
3	24.2	鮮魚	和惣菜
4	24.1	鮮魚	中華惣菜
5	23.7	鮮魚	魚卵

● サラダ油との併買ベスト5

1	24.9	サラダ油	醤油、（木曜日）
2	21.7	サラダ油	ティッシュ、（木曜日）
3	16.2	サラダ油	砂糖、（木曜日）
4	11.0	サラダ油	小麦粉
5	10.6	サラダ油	カップ麺、（木曜日）

● 豚肉との併買ベスト5

1	61.7	豚肉	牛乳、中華の素、調味料
2	59.5	豚肉	ハム、中華の素
3	58.4	豚肉	食パン、牛乳、中華の素
4	57.7	豚肉	野菜類、中華の素
5	56.7	豚肉	牛乳、中華の素

● 焼肉のたれとの併買ベスト5

1	13.9	焼肉たれ	その他精肉
2	9.2	焼肉たれ	輸入牛
3	4.0	焼肉たれ	ドレッシング
4	3.5	焼肉たれ	調味料、（日曜日）
5	3.5	焼肉たれ	国産牛

● 木曜日の併買ベスト5

1	146	サラダ油、醤油	木曜日
2	129	サラダ油、カップ麺	木曜日
3	128	サラダ油、ティッシュ	木曜日
4	125	ティッシュ、カップ麺	木曜日
5	117	ティッシュ、トイレット	木曜日

図 2　バスケット分析の例

4. 顧客分析の方法

データマイニングは，考案されて以来，マーケティングの分野で顧客の特性を分析するために最も多く利用されてきた．この分野でデータマイニングがどのように利用されるかを検討するために，最近のマーケティングにおける顧客分析の要件を考える．

最近のCRMが目指す既存顧客からの売上げを拡大するマーケティングでは，顧客の特性を明らかにし，個々の顧客を識別することが重要になる．

たとえば，販売促進のためのプロモーションを企画する場合には，どのような特性をもつ顧客がどれくらいいるかを分析し，同じ特性を持つ顧客に対して戦略を立てることになる．自宅のパソコンでメールを行っている顧客に新機種のキャンペーンを行うような場合である．このとき，「自宅のパソコンでメールを行っている顧客」を特定する必要がある．パソコンを持っていても必ずしもメールを行っているとは限らないので，パソコンを持っている顧客をさらに分類しなければならない．このように顧客分析の要件の一つは顧客のグループ分類である．

次に，この販売をテレマーケティング（TM）やDMにより行うとすれば，少しでも受注の可能性の高い顧客を選び出して電話やDMで訴求することになる．すなわち，電話をかける顧客の順番の決定やDMを送付する顧客の特定が必要である．こ

れは顧客選択の優先順位を決めるランク付けである．

このように顧客分析の基本は，顧客のグループ分類と顧客のランク付けであるといえる．

4.1 顧客のグループ分類

顧客をグループ分類する目的は，どういう特性の顧客がどのくらいいるかを把握することである．マーケティング戦略の立案や販促のためのプロモーションを考える際の基本情報となる．

顧客をグループ分類する要因として，①性別や年齢などの顧客属性，②過去の取引の回数や金額などの購入実績，③過去に購入した商品や時期などの取引内容がある．

顧客属性による分類は，分析の方法が簡単で分りやすいが，一般には，年齢，性別，住所などの基本的な情報しかデータとして得られないため詳細な分析は難しい．また，職業，年収，家族などの情報は，時系列的に変化するので注意が必要である．これらのデータを収集し最新の状態に維持するには膨大な費用を要する．また，顧客属性から顧客の本当の特性を見付けるのは困難である．

購入実績により顧客を分類するときの評価尺度としてRFMがよく用いられる．R（Recency）は最新購入日で顧客と最後に取引があった日付である．F（Frequency）は購入回数で今までの取引の回数，M（Monetary）は購入金額で今までの取引金額の合計である．これらは顧客の購買力を量的に評価することはできるが，何を買うかなど質的な評価はできない．

顧客属性や購入実績による顧客のグループ分類では，顧客の購買行動を正確に把握することは難しい．過去の取引の内容を分析して，いつ，どこで，なにを購入する顧客かを予測することが必要である．取引内容としては，いつ（シーズン，月，曜日，時間帯，…），どこで(海外，国内，県外，市内，町内，店種，個店，…），なにを(商品，サービス，オプション，支払方法，…）などを考える．表2は顧客グループ分類の例である．

表 2　利用店，利用内容による分類例と利用地域による分類例

利用パターン　特　徴

利用パターン（1）	百貨店や専門店でのショッピング中心
利用パターン（2）	ホテル等の旅行関係、キャッシング
利用パターン（3）	飲食店中心
：	：

地域パターン　特　徴

地域パターン（1）	現住所の近傍のみ
地域パターン（2）	国内のみ広く分布
地域パターン（3）	海外中心
：	：

一般に，商品分類はあっても顧客分類というものを設定している企業は少ない．顧客を評価する共通の尺度としての顧客分類が必要である．

4.2 顧客のランク付け

顧客をある基準にしたがって評価することは，マーケティングにおいては常に必要となる．たとえば，DM 等で販売促進を行う場合には，どの顧客を対象にすれば効率的かが問題になり，見込み度の高い順に顧客を選択することになる．また，商品代金の支払いや貸付金の回収などでは，与信のための評価が必要で，信用度の低い顧客の特定を行わなければならない．これらはある評価基準に従い顧客をランク付けすることである．顧客ランク付けは以下の目的で行われる．

1） 優良顧客（得意顧客）の特定…購入可能性の高い順に顧客をランク付けする．
2） 問題顧客(不正，与信)の特定…信用リスクの高い順に顧客をランク付けする．
3） 脱落顧客（解約，退会，休眠）の特定…解約・退会の危険性の高い順にランク付けする．
4） プロモーション対象顧客（DM，TM）の特定…DM や電話に反応しやすい順にランク付けする．

顧客をランク付けする最も簡単な方法は，RFM のいずれか一つの指標を用いる方法である．R がよく使われるが，最近に取引があった顧客ほど次も取引がある可能性が高いのはどの業種にも共通している．次に，R と F または R と M など二つの指標を組み合わせてランク付けの精度を上げようとする．この場合，R や F などを適当にカテゴリ化してクロス表を作ってランク付けを行う．

さらに精度を上げるためには顧客属性や取引内容など多数の要因を追加する必要があるが，要因の数が多くなるとクロス表のような集計ベースでは分析や管理が難しくなる．そこで，多数の要因を総合的に評価した数値（スコア）を与えて顧客のランク付けを行うことを考える．これをスコアリングという．スコアリングは多数の要因を考慮するためランク付けの精度に優れている．また，個々の顧客に評価値が付いているため，顧客を特定する場合など，運用上も効率的である．

5. MiningPro 21 によるデータマイニングの実現

第 3 章で述べたように顧客分析の基本は顧客のグループ分類と顧客のランク付けである．グループ分類やランク付けは，データマイニングを用いれば高い精度で効率的に行うことができる．実際のデータマイニングのソフトウェアでこれらをどう実現すれば効果的な分析が行えるかを，MiningPro 21 を例に紹介し，データマイニングのためのソフトウェアの要件を考える．MiningPro 21 は，日本ユニシスが開発したデータマイニングのためのソフトウェアで，「予測」「分類」「判別」「相関」のマイニング手法を提供している．

5.1 顧客のグループ分類と「分類」機能

顧客の購買特性を把握するためには，過去の取引内容により顧客を分類することが必要である．一方，グループ化の手法としてはクラスター分析が一般的である．ここで，取引内容として過去の購入商品から顧客をクラスター分析でグループ分類する手順を示す．

5.1.1 顧客グループ分類の手順

商品の取引内容データから顧客のグループ分類は以下のように行う．

1） 分析用データ

顧客の購入商品のデータには表 3 のように顧客別に商品別に購入有無が記録されているものとする．購入回数が多い顧客でも商品別に見れば，全ての商品に購入があることは少なく，ほとんどの商品は 0（購入無し）となっている．

表 3 顧客の商品取引データ

顧客番号	婦人	服飾	紳士	雑貨	･･･
0011130		○		○	
0016276	○	○	○		
0017007		○		○	
0120225	○				
0569872				○	
0281954			○		
0493458	○				
0678429		○		○	
1153004	○	○			
1312523			○	○	
⋮	⋮	⋮	⋮	⋮	

○：購入あり

2） クラスター分析

クラスター分析（k-means 法）は座標空間上の点を球状にグループ化する方法である（図 3）．すなわち，グループ内分散が最も小さくなるように各点（顧客）をグループ化するものである．この方法ではデータは連続量であることが必要である．表 3 の商品取引データではほとんどの値が 0 で上手く分類することはできないため，何らかの方法で連続量に置き換える必要がある．

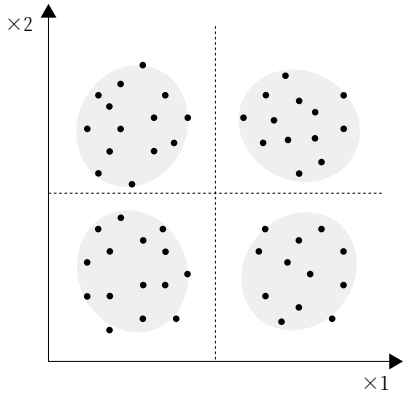


図 3 クラスター（k-means 法）の概念

3） 数量化による商品と顧客の座標配置

商品間の距離や顧客どうしの距離を考え，商品や顧客の座標配置を作成する．多くの顧客が併買している商品は共通性がある（距離が近い）と考え，座標空間

上の近くに配置する．だれも併買しない商品は共通性がない（距離が遠い）と考え，遠くに配置して商品の座標配置（商品マップ）を作成する．同様に，同じ商品を購入している顧客は共通性が高いと考え，顧客の座標配置（顧客マップ）を作成する．これらは数量化理論Ⅲ類などの方法で実現可能である．商品マップと顧客マップの例を図 4 に示す．図 4 のように連続空間上に顧客が配置できればクラスター分析（k-means 法）によるグループ化も可能になる．

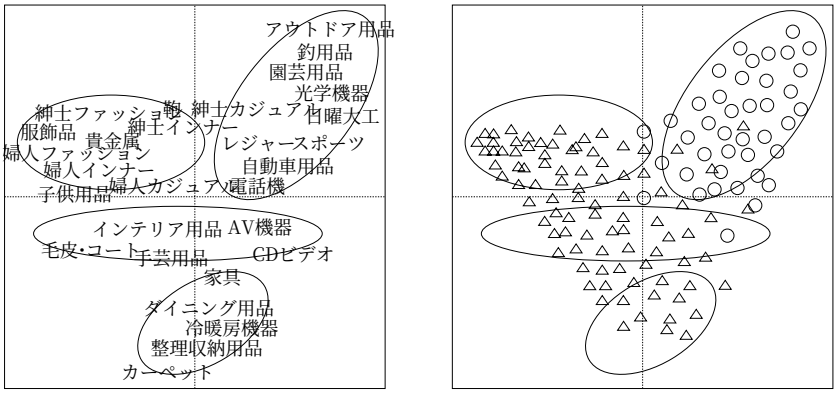


図 4　商品マップと顧客マップの例

4）　顧客グループの特徴の把握

商品マップと顧客マップの同じ位置の商品と顧客は対応しており関連があるといえる．すなわち，顧客マップ上のある顧客グループと同位置にある商品群は，そのグループの顧客群から特に購入が多い商品である．座標配置が多次元になった場合，この関連を図から読み取ることは困難である．顧客グループ毎に購入商品を集計して特徴を把握することになる．また，購入商品だけでなく顧客属性や他の要因の集計を行うことで，より詳細にグループの特性を把握することができる．

5.1.2　MiningPro 21 の「分類」機能

前節で検討したように顧客のグループ分類は以下のように行えばよいことが分る．

1）　商品マップおよび顧客マップの作成（数量化理論Ⅲ類）

2）　顧客マップ上の顧客のグループ化（クラスター分析）

3）　顧客グループの特徴の把握（顧客グループ別集計）

MiningPro 21 の「分類」機能にはこれらの手順が組み込まれており，いくつかのボタンを押すだけで実行できるようになっている．実行例を以下に示す．

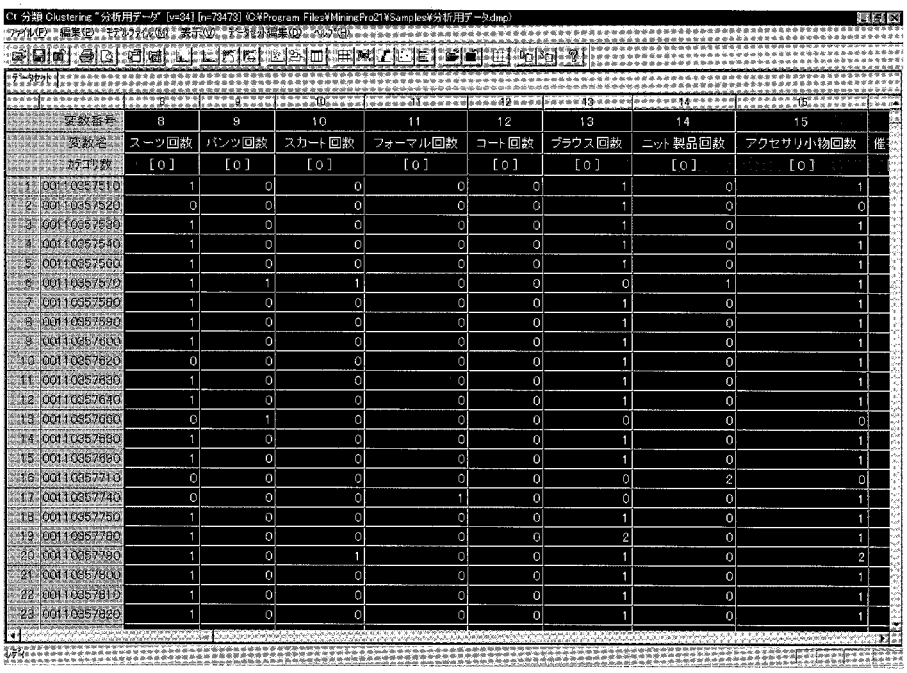
1）　変数の選択

スプレッドシート上で変数（商品）を選択する（図 5）．

2）　商品マップと顧客マップの作成

「データ読み込み」ボタンを押して，商品マップと顧客マップを作成する（図 6）．

3）　顧客のグループ化



	変数番号	8	9	10	11	12	13	14	15	
	変数名	スーツ回数	パンツ回数	スカート回数	フォーマル回数	コート回数	ブラウス回数	ニット製品回数	アクセサリ小物回数	催
	カテゴリ数	[0]	[0]	[0]	[0]	[0]	[0]	[0]	[0]	
1	0011 0357510	1	0	0	0	0	1	0	1	
2	0011 0357520	0	0	0	0	0	1	0	0	
3	0011 0357530	1	0	0	0	0	1	0	1	
4	0011 0357540	1	0	0	0	0	1	0	1	
5	0011 0357550	1	0	0	0	0	1	0	1	
6	0011 0357570	1	1	1	0	0	0	1	1	
7	0011 0357580	1	0	0	0	0	1	0	1	
8	0011 0357590	1	0	0	0	0	1	0	1	
9	0011 0357600	1	0	0	0	0	1	0	1	
10	0011 0357620	0	0	0	0	0	1	0	1	
11	0011 0357630	1	0	0	0	0	1	0	1	
12	0011 0357640	1	0	0	0	0	1	0	1	
13	0011 0357660	0	1	0	0	0	0	0	0	
14	0011 0357680	1	0	0	0	0	1	0	1	
15	0011 0357690	1	0	0	0	0	1	0	1	
16	0011 0357713	0	0	0	0	0	0	2	0	
17	0011 0357740	0	0	0	1	0	0	0	1	
18	0011 0357750	1	0	0	0	0	1	0	1	
19	0011 0357760	1	0	0	0	0	2	0	1	
20	0011 0357780	1	0	1	0	0	1	0	2	
21	0011 0357800	1	0	0	0	0	1	0	1	
22	0011 0357810	1	0	0	0	0	1	0	1	
23	0011 0357820	1	0	0	0	0	1	0	1	

図 5　MiningPro 21 の変数選択画面

「分類の作成」ボタンを押して，顧客のグループ分類を作成する（図 7）．

4）　顧客グループ別集計

「分類別に集計」ボタンを押して，顧客グループ別に各商品の平均購入回数を集計し，各グループの特徴を把握する（図 8）．

顧客マップの作成（数量化理論Ⅲ類）→顧客のグループ化（クラスター分析）→顧客グループの特徴把握（グループ別集計）が三つのボタンを押すだけで実行されるようになっている．これは，単に操作性を工夫したということではなく，分析がどのように行われるかを想定しているために実現できるものである．

結果の出力は，画面上部にタグがあり，詳細な分析結果の選択表示が可能である．しかし，通常の分析では，図に掲げた 2 種類から 3 種類くらいのグラフを見れば，統計学の面倒な知識がなくても，直感的に判断できるようになっている．

元のデータが 0・1 データまたはそれに近いデータの場合，数量化理論Ⅲ類を用い顧客マップを作成したが，元のデータが連続量の場合には主成分分析などを用いて顧客マップを作成するか，連続データから直接クラスター分析を行ってもよい．

5.2　顧客のランク付けと「予測」機能

顧客のランク付けはスコアリングによる方法が精度も良く，顧客個人を特定できることから運用上も利点が多い．スコアリングは回帰分析やニューラルネットなどで行うことができる．ニューラルネットは，パラメータの推定が面倒でロバスト性に欠けることや，大量サンプルが扱えないことなどから，回帰分析で行う方が安定した解を得られることが多い．ここでは回帰分析によるスコアリングの方法を検討する．

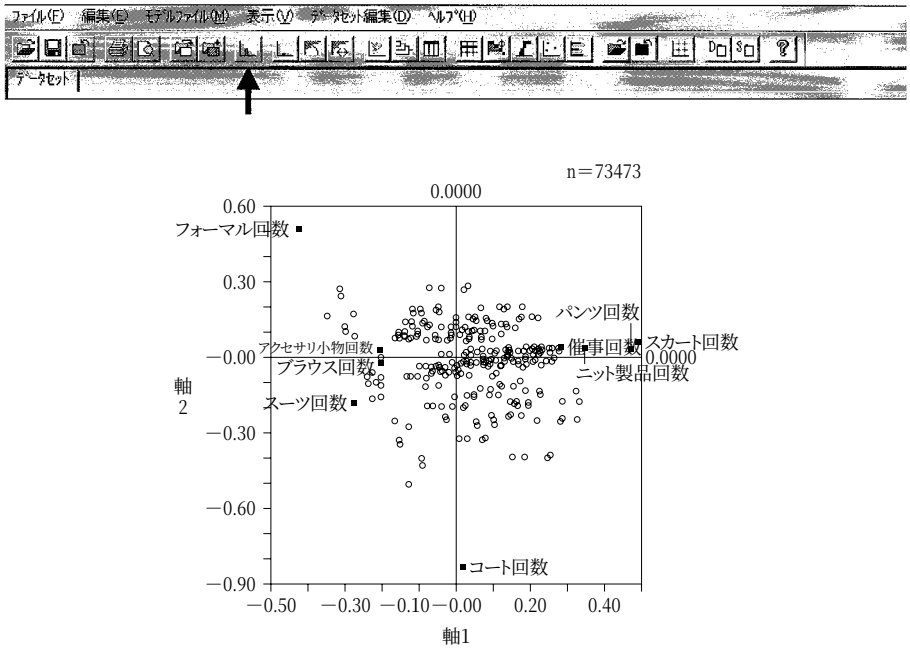


図 6　商品マップと顧客マップの作成

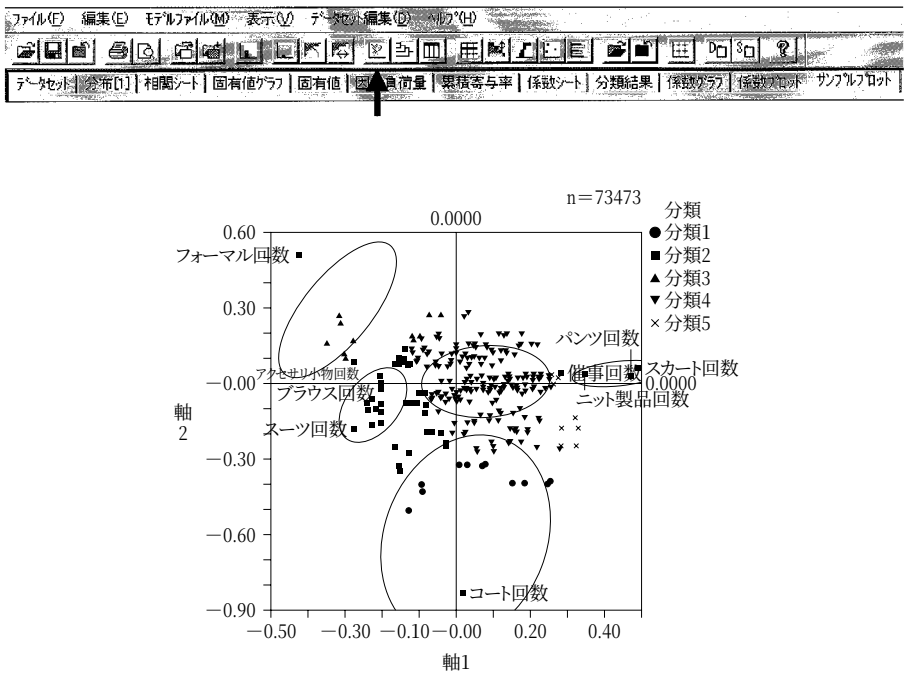


図 7　顧客のグループ化

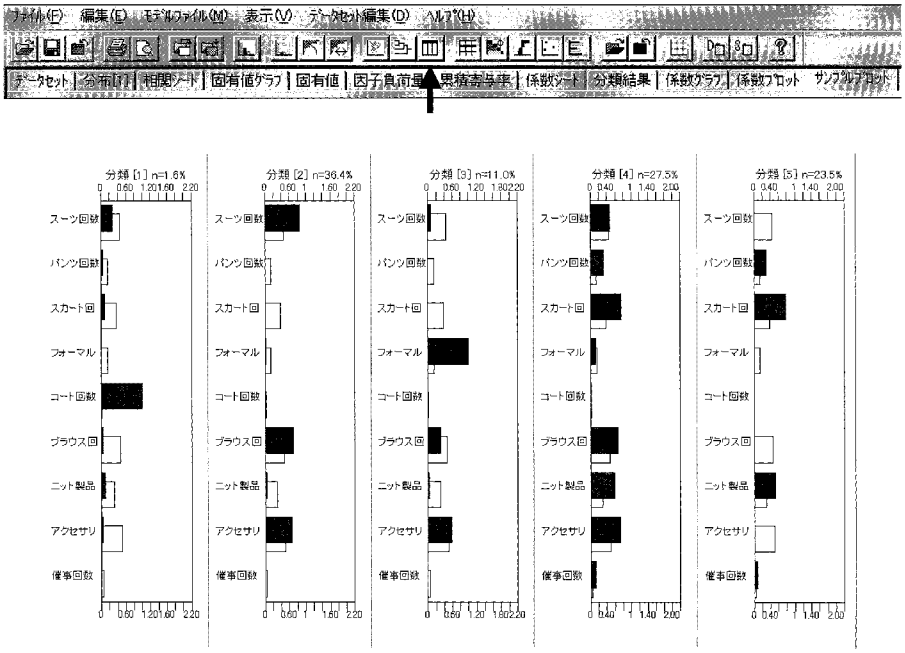


図 8　顧客グループ別集計

5.2.1　顧客ランク付けの方法

あるキャンペーン期間中の各顧客の購入金額を回帰分析により予測することを考える．説明変数は顧客属性や RFM などキャンペーン開始前の顧客の状態を表す変数である．分析結果の予測値と実績値の関係を図 9 に示す．図は 5 万件のサンプルを使って計算している．重相関係数は 0.203 で統計学的には大きな値ではない．一般に，個人顧客のデータを使ってこのような回帰分析を行った場合，重相関係数は 0.2 から 0.3 くらいの値になるのが普通で，一見分析結果は役に立たないように見える．

顧客を予測値の大きい順に並べ，人数が等しくなるように 10% ずつの 10 ランクに分類する．ランク別の反応率（購入金額が 1 以上の割合），平均購入金額，合計購入金額を表 4 に示す．表 4 の顧客ランクは下位のランクは上位の顧客を含む累積となっている．

図 9 では予測値と実績値は無相関に見えたが，表 4 では上位ランクほど反応率や平均購入金額は高く，このランクは顧客の購入を予測するのに有効であることが分る．さらに，上位 10% の顧客を見ると，反応率は全体平均の 3.18 倍，平均購入金額は 3.98 倍になっている．上位 50% でも，反応率は 1.65 倍，平均購入金額は 1.79 倍である．また，合計購入金額では，上位 10% の顧客で全体売上げの 39.8%，上位 50% の顧客で 89.5% の売上げが確保されることが分る．統計学的には有意でない結果でも実用上は有効な結果が得られている．

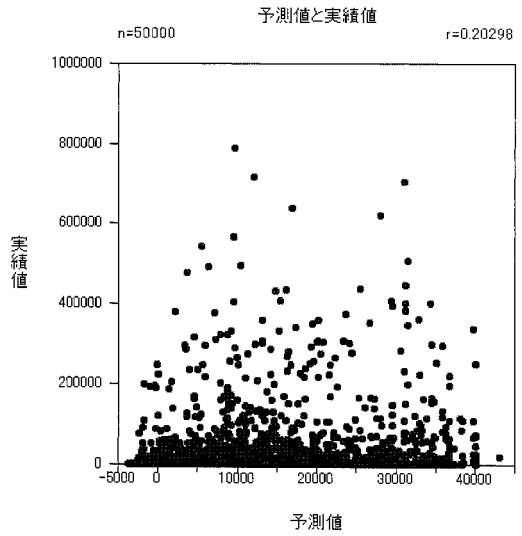


図 9　予測値と実績値の関係

表 4　顧客ランク（累積）別反応率，平均購入金額，合計購入金額

	顧客ランク	顧客数	反応率	倍率	平均購入金額	倍率	合計購入金額	構成比
1	ランク[01]10%	5,000	39.0%	3.18	27,536	3.98	137,680,191	39.8%
2	ランク[02]20%	10,000	31.5%	2.56	20,726	3.00	207,255,668	60.0%
3	ランク[03]30%	15,000	26.6%	2.17	17,267	2.50	258,999,486	75.0%
4	ランク[04]39%	19,999	23.0%	1.88	14,479	2.10	289,560,987	83.8%
5	ランク[05]50%	25,004	20.2%	1.65	12,365	1.79	309,184,012	89.5%
6	ランク[06]59%	29,992	17.9%	1.46	10,771	1.56	323,043,979	93.5%
7	ランク[07]69%	34,987	16.1%	1.31	9,444	1.37	330,414,679	95.6%
8	ランク[08]79%	39,999	14.6%	1.19	8,419	1.22	336,766,424	97.5%
9	ランク[09]89%	44,998	13.3%	1.08	7,604	1.10	342,165,611	99.0%
10	ランク[10]100%	50,000	12.3%	1.00	6,911	1.00	345,538,810	100.0%

5.2.2　MiningPro 21 の「予測」機能

前節のような分析は，MiningPro 21 の「予測」機能では，モデル作成画面とモデル検証画面の二つの画面で行うことができる．

図 10 はモデル作成画面である．推定結果である各カテゴリのウェイト（0 を中心に左右に棒が出ているグラフ）やモデルに取り込む前の変数の状態（左から棒が出ているグラフ）が表示されている．変数の選択は各グラフを直接マウスでクリック（黒い影があるグラフ）して行う．変数の有効性は棒グラフにより確認できる．

図 11 はモデルの検証画面である．表 4 で示した反応率，平均購入金額，合計購入金額がグラフで示されている．モデルの有効性を統計学的にではなく実用上の見地から視覚的に確認できるようになっている．モデルを検証する方法は多様だが，この図はスコアリングの結果を検証する一つの方法を示している．

以上のように二つの画面を見てモデルを作成しその検証を行う．そのとき，結果を各種の統計量で出力するのではなく，実際のデータを集計して見せることで，統計学の専門的な知識が無くとも常識的な判断のみで分析が進められるようになっている．

さらに，分析の結果は各サンプルに反映され，顧客毎にスコアが自動的に付与される．スコアはこれも自動的に確率や金額に換算される．

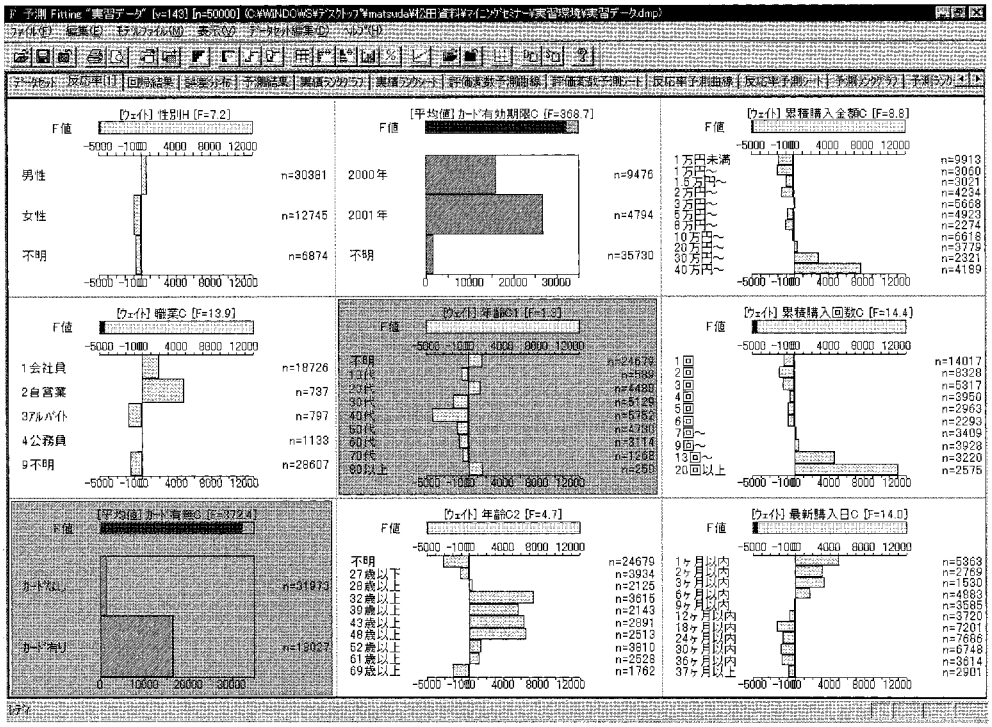


図 10　モデル作成画面

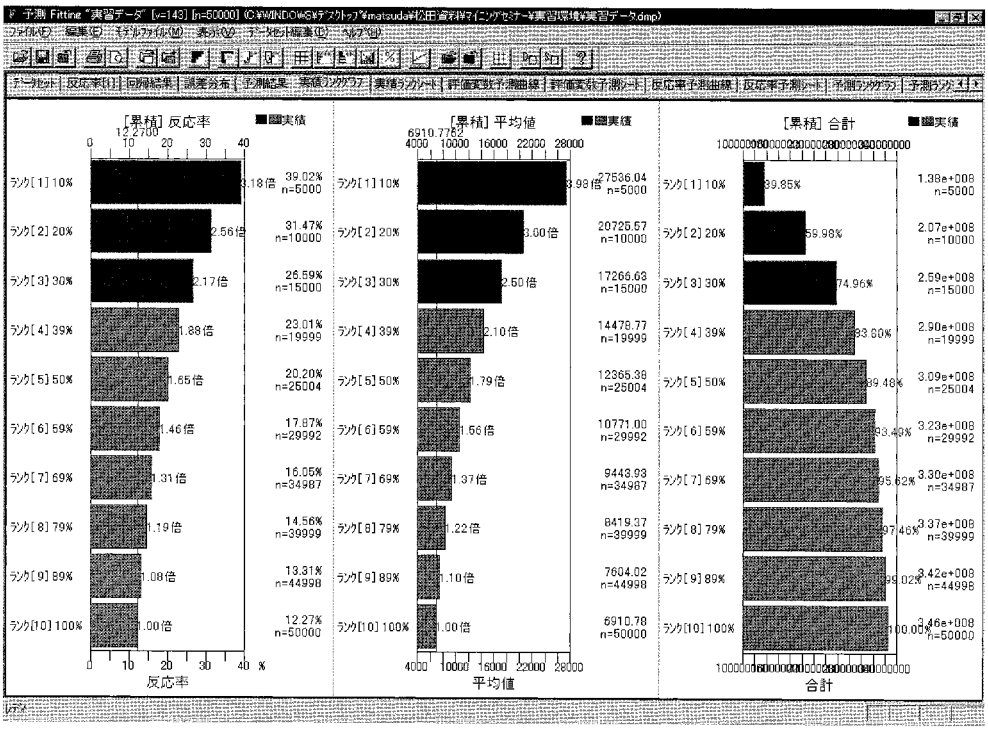


図 11　モデル検証画面

5.3 MiningPro 21 の特徴

MiningPro 21 の特徴を整理し，マーケティングなどの実務で利用するデータマイニング用ソフトウェアに必要な条件を考える．

1） 分析の手順がシステム化されている

「分類」では，顧客マップの作成（数量化理論Ⅲ類）→顧客のグループ化（クラスター分析）→顧客グループの特徴把握（グループ別集計）という分析の手順を三つのボタンを押すだけで実行できた．「予測」でも二つの画面でモデルの作成と検証を行うと，自動的に確率や金額という尺度でスコアが各顧客に付与された．一般的にデータ解析は，①モデルの作成，②パラメータの推定，③モデルの検証，④予測という手順で行われるが，2，3 のボタン操作と 2，3 の画面を見るだけで自動的に実行されるようになっている．

2） 分析のモデルが組み込まれている

顧客分析の基本である顧客のグループ分類や顧客のランク付けは決まったやり方があるわけではなく，どのような方法でも実現可能である．しかし，統計学などの専門的知識を有することと分析モデルを作成することとは別で，誰でも有効なモデルを作成できるとは限らない．MiningPro 21 には，サンプルを座標空間上に配置してグルーピングする方法，スコアリングした結果をランク分けして評価する方法など，有効と思われるモデルが組み込まれている．これらは過去のデータベースマーケティングで実践された数多くの実績に基づいている．これにより分析の経験が少ない人や分析の方法が分らない人でも過去の有益なモデルを利用することができるるようになる．

3） 専門的な知識を必要としない

前章で見たようにモデルの評価は全てグラフで行えるようになっている．これらのグラフは統計学上の数値（統計量）を表示したものではなく，実務に適用したときにどういう効果が見込めるかという見地から作成されている．したがって，専門的な知識を持たなくとも常識的な判断からモデルの評価を行うことができる．

従来，データ解析のソフトウェアは実際に使われる問題を規定せずに手法を中心に開発されてきた．「分類」や「予測」に利用可能な手法は用意されていたが，それらをどう使うかは利用者の判断であった．その結果，いくつもの高度な手法が独立に用意され一般の実務家には使いこなすどころか内容の理解も難しいものになっている．

顧客をマスで捕らえるマーケティングから顧客個人を識別するマーケティングへとマーケティングの方針が変わり，CRM が実践されている環境において，データマイニングを初めとするデータ解析は必須のものとなっている．データ解析のツールも手法を用意するだけではなく，分析の手順やモデルを包含するものが必要である．データ解析の技術を大衆化して，分析が行える人の数を増やし，その生産性を上げるようなソフトウェアが必要である．MiningPro 21 はその一つの方向を示している．

6. データマイニングの今後

インターネットなどを利用したビジネスの普及により，データマイニングの利用方

法や分析の対象に変化がおきている．

一つは，インターネットに接続してきた顧客の特性をリアルタイムに識別し，顧客の特性に合ったページやコンテンツを表示するオンライン・マイニングの必要性である．

もう一つは，文字データを分析する文書マイニングである．今までのデータマイニングの対象は数値データであったが，顧客からの質問や苦情がデジタル化された文書として保存されると，この文書を分析することが必要になる．

ここでは，今後のデータマイニングで必要となる，オンライン・マイニングと文書マイニングについて述べる．

6.1 オンライン・マイニング

従来のマーケティングでは，顧客への訴求は DM や電話などの方法で行われてきた．訴求用の顧客リストは前日にバッチ処理で作成し，そのための分析も事前に十分な時間を掛けて行うことができた. 現在はインターネットによるビジネスが普及して，顧客は Web をとおして商品を購入したりサービスを受けたりすることができる．このような状況では，顧客の状況も Web にアクセスする都度変化し，事前に分析しておくことは困難である．顧客がホームページに接続すると同時にその顧客の特性を識別し，それぞれの特性に合ったページを表示したり商品やサービスの提示を行うことが必要である．また，ホームページやコンテンツを閲覧したり，商品やサービスを購入するたびにその顧客の状況は変化しており，リアルタイムに識別情報も更新する必要がある．すなわち，データマイニングもオンラインで行う必要が出てきた．

ここで識別しなければならない顧客の特性は以下の 2 種類である．一つは，その顧客がどういう分野に興味を持っているか，どのような商品やサービスを求めているかなどの顧客の選好分野を識別することである．たとえば，航空会社の顧客の例を考える．航空会社の顧客の特性には，フライト，ツアー，グルメなど，どの分野に興味があるかの特性，目的地や利用時間帯の特性，ホームページのコンテンツ内容の特性がある．これらの特性は図 12 のように数値化しベクトル形式で各顧客に付与しておく必要がある．これを顧客のプロフィールと呼ぶ.

もう一つの特性は，顧客がその会社にとって有益かどうか期待の程度を表す特性である．たとえば，航空券を予約してくれるかどうか，他の商品を買ってくれるかどうか，キャンセルや解約の恐れはないかなどの度合いを数量化して表す（図 13）．これを顧客の優良度と呼ぶ.

顧客プロフィールも顧客の優良度もデータマイニングによって求めることができる．顧客プロフィールは「分類」の手法で求めることができる．職業や年齢などの顧客属性をいくら分析しても図 12 のような顧客プロフィールを求めることはできない．過去の取引の履歴や閲覧したホームページやコンテンツの履歴を詳細に分析することにより可能になる．取引履歴や閲覧履歴のデータから顧客グループを作成し，そのグループにどれだけ近いかを数量化しベクトル表現する.

顧客の優良度は「予測」の手法で求めることができる．購入の期待度や解約の危険度を過去の実績からスコアリングにより評価する.

顧客のプロフィールや優良度が分れば，ホームページに接続した時点でどのような

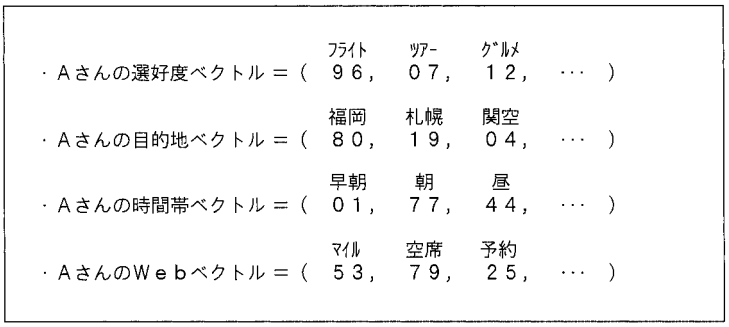


図 12　顧客プロフィール

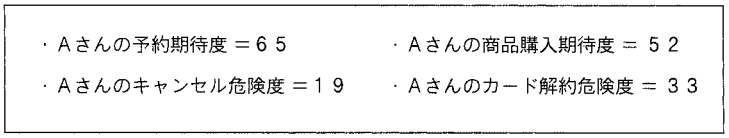


図 13　顧客の優良度

特性の顧客かが識別可能なので，その顧客に合ったページやコンテンツを提示することが可能になる．

DM やカタログなどと比べれば，ホームページでは限られた数の商品しか陳列することはできない．また，閲覧している時間も限られている．インターネットのビジネスでは，アクセスしてきた顧客を速やかに認識し，特性に合った内容を提示することが必要となる．顧客の特性を的確に識別するためにはデータマイニングは必須である．

顧客がホームページをアクセスしている間は閲覧するコンテンツが変わり，購入や解約などの取引も進んでいく．この変化に伴い顧客プロフィールや優良度もリアルタイムに更新していく必要がある．これに対応してデータマイニングもトランザクション処理が必要になる．

通常，データマイニングのソフトウェアは，分析用のデータを使ってモデルを作成し，そのデータや他のデータセットに対しまとめてスコアリングを行って各顧客に識別値を付けていた．インターネットのビジネスに対しては，新しいトランザクションデータが入力された時点で，モデルを更新し，顧客プロフィールや優良度をそのトランザクションに対して即時に計算する機能が必要になる．

6.2 文書マイニング

インターネットやコールセンターなどの新しいビジネス形態の普及にともない，顧客の問い合わせ，要望，苦情などが紙ではなくデジタル化された文書として蓄積されている．各企業ではこれらの顧客の声を反映させて商品の企画やマーケティングへの適用を行おうとしている．ある化粧品会社では，寄せられた文書は，「要望」「質問」「苦情」「感想」のいずれかのカテゴリに仕分けされる．さらに，「肌」に関する文書か，「髪」に関する文書か，「使用法」に関する文書か，「品質」に関する文書かなど，内容により分類される．これらは手作業で行われており省力化が求められている．

このように文書を内容により自動的に分類する技術と，文書をあらかじめ決められたカテゴリに自動的に判別する文書に対するデータマイニング技術が求められている．

6.2.1 文書の自動分類

文書を構成している単語の組み合わせにより文書を自動的に分類することを考える．文書の分類は顧客のグループ分類と同じ考え方で行うことができる．すなわち，単語と文書の反応表から数量化理論Ⅲ類の考え方で単語マップと文書マップを作成して文書の座標空間上の配置を求める．この空間配置上の文書はクラスター分析などでグループ化することができる．

この方法の特徴は，最初に分類の基準を作らないところにある．単語の共起関係だけでグループ分類を行い，分類してから各グループの意味付けを行う．分類の基準を作るためには全体の文書の内容を理解している必要があるが，どのような内容を含んでいるか不明な場合にもこの方法は適用可能である．また，使用する単語の組み合わせを変えれば，異なった概念の分類を作成することもできる．

分類の結果は各グループに属する帰属度で表される．一つの文書が二つの内容を持つ場合や意味があいまいな場合の対応も可能である．

6.2.2 文書の自動判別

文書をあらかじめ決められた「要望」「質問」「苦情」「感想」のどのカテゴリに属するか判別するために四つの判別関数を作成する．すなわち，「要望」か否かを判別する関数，「質問」か否かを判別する関数，「苦情」か否かを判別する関数，「感想」か否かを判別する関数である．これにより四つの判別確率 $P_{「要望」}$，$P_{「質問」}$，$P_{「苦情」}$，$P_{「感想」}$ が計算され，これらの確率の大きさによりどのカテゴリに属するかを判別する．

人間が書いた文書は単純ではなくいろいろな要素を含む．これを単に「要望」「質問」「苦情」「感想」のどれかに仕分けすることには無理がある．四つの確率で各カテゴリへの帰属度を表現することにより，各文書が「要望」「質問」「苦情」「感想」のどの要素をどれくらい含むかが分り，択一ではない適切な判別を行うことができる．たとえば図 14 では，「苦情の要素を含む要望」とか「苦情に近い要望」と判断することが可能である．

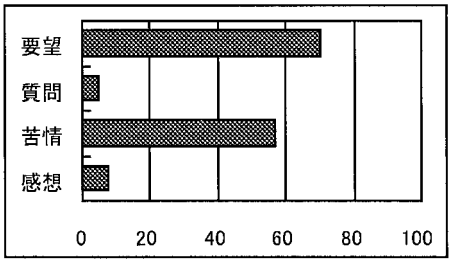


図 14　判別確率の例

6.2.3 文書マイニングの利用

以上の文書マイニングの技術により，従来のキーワードだけによる検索の他に，概

念検索，類似検索，自然文検索など多様な検索が行えるようになる．

1) 概 念 検 索

文書は「肌」「髪」「使用法」「品質」などに分類され，それぞれのグループへの帰属度が付けられている．たとえば，「品質」に関する文書を検索したいときは，「品質」への帰属度の高い文書から順に抽出すればよいことになる．さらに，「要望」「質問」「苦情」「感想」との組み合わせで『「品質」に関する「苦情」』などという文書も簡単に検索可能である．このようにキーワードではなく概念による検索が可能になる．

2) 類 似 検 索

キーワードによる検索の場合はキーワードが一致しないと文書を抽出することができない．文書分類の場合，各単語も座標空間上に配置され，単語間の距離も計算することができる．したがって，指定されたキーワードが一致しなくても，そのキーワードと最も近い単語を持つ文書を検索することができる．これを類似検索という．

3) 自然文検索

通常の日本語で書かれた文書を入力して検索を行うことも可能である．検索文のから単語を抽出して類似検索を行うことができる．顧客からの問い合わせをそのまま入力して回答を引き出すことも可能になる．

以上のような検索技術を使って，コールセンターでは寄せられた質問に対し過去の事例などを検索して回答することができる．E メールなどで寄せられた文書に対しては自動回答も可能である．商品開発部門は，商品に対する苦情や要望を解析して商品の改良や企画などを行う．マーケティング部門では，顧客の声と過去の購入実績を分析して見込み顧客の抽出を行う．企画部門では，アンケートの自由回答文を分析してさまざまな企画立案を行うことができる．文書マイニングにより顧客とのリレーションの持ち方も変化する．

7. おわりに

従来の統計的データ解析は，仮説の設定やモデル化の作業が必要であり，専門的な知識を要した．ルールの自動抽出を目的とするデータマイニングにより，データ解析がマーケティングを中心とした業務の中に取り入れられ普及してきた．しかし，データマイニング・ツールは分析手法だけを提供するものであり，実際の問題への適用の方法は示してくれない．そのため，経験の少ない人には難しいものになっている．今後は，分析の手順や適用のモデルなどを包含したデータマイニングのソフトウェアも必要となる．

インターネットによるビジネスでは，接続してきた顧客の識別はさらに重要になり，データマイニングもリアルタイムで行う必要がある．顧客の声を中心とする文書データが蓄積されると，今まで数値データだけを対象としてきたデータマイニングも文字データを扱う必要が出てきた．データマイニングも新しい分野への活用が期待されている．

参考文献
［1］ 荒川圭基（1991）: データベース・マーケティングの戦略と戦術, ダイヤモンド社.
［2］ Don Peppers, Martha Rogers（1995）ONE to ONE マーケティング, ダイヤモンド社.
［3］ Arthur M. Hughes（1999）: 顧客生涯価値のデータベース・マーケティング, ダイヤモンド社.
［4］ 沼尾他（1997）: 特集「大規模データベースからの知識獲得」, 人工知能学会誌 1997 年 7 月.
［5］ Pieter Adriaans, Dolf Zantinge（1998）: データマイニング, 共立出版.
［6］ J. A. Berry, Gordon Linoff（1999）: データマイニング手法, 海文堂出版.
［7］ ユニシスニュース 2000 年 7 月 1 日第 71 号「データマイニング特集」, 日本ユニシス.
［8］ M. J. A Berry, G Linoff（1997）: Data Mining Techniques, Willey.
［9］ G. Robert（1998）: DATA MINING, Prentice Hall.
［10］ J. R. Quinlan（1993）: C 4.5 : Programs for Machine learning, Morgan Kaufmann.
［11］ Ovum Ltd.（1997）: Ovum Evaluates : Data Minig, Ovum.
［12］ 奥野忠一, 久米均, 芳賀敏郎, 吉沢正（1981）: 多変量解析法, 日科技連出版社.
［13］ 田中豊, 脇本和昌（1983）: 多変量統計解析法, 現代数学社.
［14］ 駒沢勉（1985）: 数量化理論とデータ処理, 朝倉書店.
［15］ 飯沼一元（1989）: ニューロコンピュータ, 技術評論社.
［16］ Michael R. Anderberg（1988）: クラスター分析とその応用, 内田老鶴圃.
［17］ H. C. Romesburg（1989）: CLUSTER ANALYSIS FOR RESEACHERS, Robert E. Krieger.
［18］ T. Kohonen（1995）: Self-Organizing Maps, Springer.
［19］ H. Ritter., T. Martinetz, K. Schulten（1992）: Neural Computation and Self-Organizing Maps, Addison-Wesley.
［20］ 赤沼昭太郎, 津田宏治（2000）: サポートベクターマシン基本的仕組みと最近の発展, 数理科学 2000 年 6 月号, No.444.

執筆者紹介 松 田 芳 雄（Yoshio Matuda）
1974 年慶応義塾大学工学部管理工学科卒業．同年日本ユニシス(株)入社．オペレーションズ・リサーチ，統計解析関係のシステム開発に従事．現在研究開発室に所属．日本オペレーションズ・リサーチ学会，日本シミュレーション学会，情報処理学会会員．