マーケティング・サイエンスIII

顧客ターゲティング分析：データマイニング手法の活用

佐藤　栄作

1. はじめに

データマイニングには，目的志向的データマイニングと探索的データマイニングがある．Berry and Linoff[1]によれば，前者は，実務上の課題・目的が明確な場合に行われるものであり，データに基づき対象を分類・推定・予測することが多い．一方，後者は，データに含まれる全ての変数間の，何らかの関係性を構築することに主眼が置かれるものであり，アソシエーションルールないし関連事象グループ化，クラスタリング，記述とビジュアル化がその主な領域となる．ただしBerry and Linoffが指摘しているように，両者は補完的なものであり，探索的データマイニングを行うことで得られたルール・知識は，目的志向的データマイニングにおける予測モデルの精度を向上させるために役立てることができる．実際のデータマイニングの作業では，このような双方向的な作業を繰り返しながら，最終的な目標に向かって作業を進めることとなる．本稿では，データマイニングの手法を利用して，顧客ターゲティングを行う方法について概説する．ただし，ここではデータマイニング手法の解説を行うことが目的ではなく，あくまでもマーケティングにおけるターゲティングへのデータマイニング手法の適用に主眼を置いているため，目的志向的データマイニングを中心に，事例を交えながら説明することとする．データマイニングの各手法についての解説は，Berry and Linoff[2]やMitchell[3]を参照いただきたい．次節では，セグメンテーションとターゲティングに関する既存研究について若干のレビューを行う．節3では，顧客ターゲティングのためのデータマイニング・プロセスについて概観し，節4において事例を提示する．最後に節5でまとめと考察を行う．

さとう　えいさく
(財)流通経済研究所
〒141-0031 品川区西五反田7-23-1

2. セグメンテーションとターゲティング

マーケティング計画を立案する上で，対象とする市場を異質な消費者の集合として捉え，それらを同質なサブグループに分け，いずれかのサブグループをターゲットとしてマーケティング活動を展開するというような，Smith[4]によって提唱されたマーケットセグメンテーションのコンセプトが，マーケティング研究および実務の双方において基本的な要素の一つとなっている．そしてこのコンセプトは，製品差別化から広告，価格，流通チャネルまでマーケティング戦略全般の課題解決のために広く用いられている．これに伴って，目的に応じた様々なセグメンテーション手法が提案されてきた[5, 6]．

さらに情報技術の発展とコストの低下を背景として，個々の顧客に関する属性や行動データが大量に蓄積可能となり，加えてそれらの顧客に対する直接的なアプローチが可能となってくるに従い，ダイレクト・マーケティングやデータベース・マーケティングの可能性が高まり，より小規模で同質性の高いセグメントの抽出が求められるようになってきた[5]．このような要請に対してBult and Wansbeek[7]は，ダイレクト・メールのターゲット選定手法について研究し，限界費用と限界収益を考慮することによって，RFMモデルなど既存手法よりも効率的なターゲット選定を可能とする手法の提案を行っている．また，Rossiら[8]は，既存のダイレクト・マーケティングにおけるターゲティングのための手法が，一つのマーケティング・アクションに関する顧客反応のスコアリングに主眼を置いていることを批判的にとらえ，これに対して顧客あるいは顧客セグメントごとにマーケティング・アクションをカスタマイズすることで，さらに収益性を向上できると主張し，そのための手法の提案と実証分析を行っている．加えてRossiらは，マーケティング担当者が，顧客母集団のブランド選好や価格感度の分布のみが利用可能な状況から，個々の顧客の購買履歴データ

とコーザルデータ（売上変動要因を説明するデータ）が利用可能な状況までを想定し，それらの情報レベルに応じて提案手法を適用し収益性の評価を行った．その結果に基づき，たとえ短期間の購買履歴しか利用できないような状況でも，ダイレクト・マーケティング活動の収益性を改善できる可能性のあることを示した．Long and Schiffman[9]も，航空業界における事例を取り上げ顧客セグメントに関する検討を行った上で，Rossiらと同様に顧客への提示内容のカスタマイズの必要性を主張している．

このようにマーケティング研究領域，特にダイレクト・マーケティングあるいはロイヤルティ・マーケティングの研究では，注力すべき顧客を特定する（ターゲティングの）ためのスコアリング手法の開発から，顧客セグメンテーションと情報提供や商品提案（以下，オファーと呼ぶ）のカスタマイズへと視点が移りつつあることが伺える．

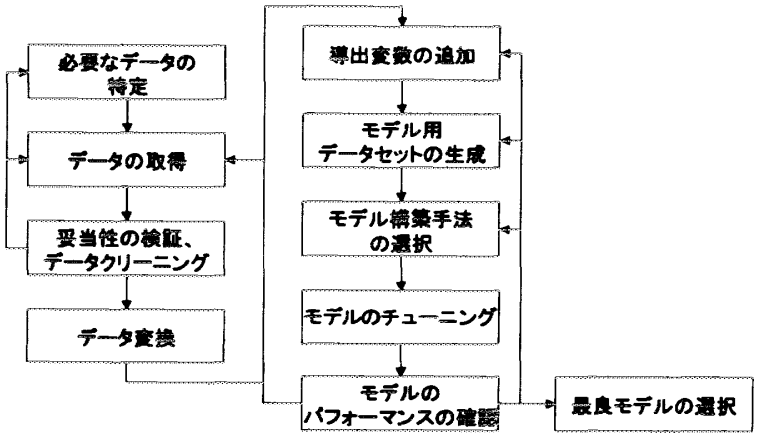
3. 顧客ターゲティング分析の流れ

前節では，マーケティング活動におけるターゲティングに焦点を当て，これまでの研究の流れを簡単にまとめた．ここでは特にダイレクト・マーケティングあるいはロイヤルティ・マーケティングの観点から，顧客ターゲティングについて論ずることとする．つまり，あるオファーに対する反応を，顧客あるいは顧客セグメントごとのスコアとして表現し，コスト制約に基づき，スコアの高い方から当該オファーに対するターゲット選定を行うための顧客ターゲティングを中心として扱う．

このような顧客ターゲティングのためのスコア算出を行うには，冒頭に述べた目的志向的データマイニングのアプローチをとることが一つの方向である．具体的には，顧客の属性や購買履歴データを用いて，特定のオファーに対する反応確率を予測するモデルを構築していくことが中心となる．本節では，このプロセスについてBerry and Linoff[1]を参考にしながら概説する．

図1は，データマイニングにおける予測モデル構築プロセスを示している．全体として10個のステップから構成されている．第1ステップは「必要なデータの特定」である．目的に適合し，入手可能なデータを特定する．同時に，過去の反応に対する観測結果のデータも当然のこととして必要となる．第2ステップでは，特定したデータを実際に取得する．続いて第3ステップでは，取得したデータの各フィールドに関して，欠測値や異常値，論理的矛盾の有無の検証を行い，異常がある場合にはその修正を行う．第4ステップでは，目的に対して取得したデータの集計レベルを合わせるなどの変換を行う．第5ステップでは，取得したデータから予測に必要と想定される導出変数を作成する．例えば，取得したデータがトランザクション・データであれば，そこから1ヶ月の来店回数や総購買金額などが導出変数として作成されるかもしれない．ここまでの段階で，予測モデル構築のために必要なデータの準備が一応完了する．

第6ステップ以降がモデル構築のステップである．第6ステップでは，準備したデータを学習・検証用，評価用のデータセットに分割する．第7ステップでは，予測モデルを構築するための手法を選択する．予測を行うと同時に，顧客の反応メカニズムについての解釈も必要とする場合は決定木が選択され，その他の場合には決定木以外にニューラルネットなどの手法も利用



出所：Berry and Linoff[1]pp. 48 Fig. 3.3を加筆修正

図1　予測モデル構築プロセス

される．第8ステップでは，学習用データセットを利用してモデルの学習が行われ，得られたモデルを検証用データセットに適用して精度分析を行いながら，モデルのチューニングが行われる．第9ステップでは，第8ステップで得たモデルを，学習や検証に利用していない評価用データセットに適用し，最終的なモデルのパフォーマンスを確認する．この結果，十分な精度が得られていると判断できれば，そのモデルを最終的なモデルとして採択し，実際のターゲティングに利用することとなる．また，複数の手法を用いて複数の予測モデルを構築している場合には，最終的にそれらのパフォーマンスを比較して，最良モデルを選択する．

このような流れに沿って顧客ターゲティングのための予測モデルを構築することとなるが，図1の各ステップは逐次的に進むことができるとは限らない．図1中の矢印に示されるように，それぞれのステップにおいて問題があれば，問題に関連すると想定される前のステップに立ち戻り，改良を加えて以降のステップを再度行うといった反復的なプロセスとなる場合が多いことに留意されたい．

4. 顧客ターゲティングのための分析事例

前節で示した顧客ターゲティングのための予測モデル構築プロセスに基づき，ここでは架空の事例を設定して説明を行う．ただし，導出変数の追加以前のステップは，状況に依存する部分が多いため立ち入らず，主にモデルの学習・評価に関する部分を中心に説明することとする．なお，ここで設定した事例は，Berry and Linoff[1] (pp. 421-423) を参考に構成したものである．

4.1 分析事例の設定

初めに筆者が設定した事例の概要を説明する．フリークエント・ショッパーズ・プログラム (FSP) を実施している某スーパーマーケットが，ヨーグルトの購買が少ない顧客の中から潜在的なヨーグルト愛好者を特定し，それらの顧客に自店舗でのヨーグルト購買を促進させるオファーを提示したいと考えている．具体的には，携帯メールやKIOSK端末による特定顧客に限定したクーポン提供や，ポイントカードのポイント付与などのプロモーションを想定している．このようなことを行うためにこのスーパーマーケットでは，顧客の属性や店舗利用状況などから，潜在的ヨーグルト愛好者を特定する分析モデルを構築することとした．それに際して，このスーパーマーケットでは，1ヶ月のヨーグルト購買金額が全体の上位30%に入り，かつ食品に占めるヨーグルト購買金額構成比が全体の上位30%に入る顧客を，ヨーグルト愛好者と定義することとした．さらに，性別と年齢をもとにカテゴライズされた性年齢区分，家族人数，65歳以上の同居家族の有無，6歳未満の幼児の有無，小学生の有無，13～19歳の子供の有無，店舗までの来店手段，5分刻みの来店時間区分などカード会員のアンケートで補足していた項目を，顧客属性データとして利用することとした．これと合わせて，過去の購買履歴データから，月の休日来店比率を算出し変数に加えることとした．以上がここで扱う事例の設定である．

4.2 データの準備

上記の設定に対応するデータを，実際のスーパーマーケットのデータから作成し以降の分析を進めることとした．利用したデータは，関東のあるスーパーマーケットのFSPデータである．全カード会員顧客の中からランダムサンプリングを行った上で，さらにその中からヨーグルト購買を行った顧客のみを対象とし，最終的に460人分の購買履歴データを抽出した．購買履歴データの抽出期間は，2001年4月から6月までである．このデータに関して，上記の定義に従い，ヨーグルト愛好者とそれ以外を区別する変数を追加した．

なお，ここでは実際に携帯メール等のプロモーションを行ってはいないので，上記460人の顧客データを学習用・検証用・評価用に分割し，評価用データに基づく精度分析結果の提示までを説明の対象とする．

対象460人の顧客について以下の表1のように分割を行った．ただし，学習用・検証用についてはモデルの学習段階で交差妥当化を行う際に複数に分割される

表1 顧客データの分割

	ヨーグルト愛好者	その他	計
学習・検証用	67 14.6%	360 78.3%	427 92.8%
評価用	6 1.3%	27 5.9%	33 7.2%
計	73 15.9%	387 84.1%	460 100.0%

ため，ここでは学習用・検証用と評価用の二つの分割結果を示す．

4.3 モデルのチューニング

表1では，ヨーグルト愛好者と設定された顧客数とその他の顧客数が極端に異なり，アンバランスな状況となっている．このような状況のデータに決定木などの手法を適用し学習させると，全ての顧客をヨーグルト愛好者以外と予測するルールが抽出されることが度々起こる．しかしこのようなルールが抽出されても，ヨーグルト愛好者を特定するという当初の目的には意味を成さないモデルとなってしまう．このケースでも上記の学習・検証用データを使って決定木の学習を行ったところ，そのようなルールが抽出された．

このような無意味なモデルとなることを回避する方法として主に二つの対策を講じることができる．一つは，オーバーサンプリングである．これは，データ数の少ない方（この場合はヨーグルト愛好者）に関して，重複を許容してサンプリングを行い，擬似的にデータ数を増加させるというものである．ただし，重複して抽出するデータの変数に誤りが含まれている場合，それが増幅されてしまうため，データ数が十分に確保できている場合には，むしろデータ数が多い方からランダムサンプリングを行い減少させた方がよいとの指摘もある[1]．もう一つの方法は，ヨーグルト愛好者をそれ以外と判断したり，その逆など，誤分類に対する重みを変更することである．この事例では，ヨーグルト愛好者を特定したいのであるから，ヨーグルト愛好者をその他と判断する誤りに対する重みを，その逆の誤りよりも大きく設定し，それによってヨーグルト愛好者については可能な限り正しく分類させるように調整することとなる．ただしこの場合，ヨーグルト愛好者でない顧客をヨーグルト愛好者と判断する誤りについては大目に見るということになるため，その部分に関する精度は低下することとなる．

本事例のデータについてもオーバーサンプリングによりデータ数を擬似的に増加させた上で，決定木をデータに適用した．モデル構築に際しては，決定木の剪定度および誤分類に対する重みを変化させながら，学習および交差検証を交互に実施し，モデルの調整を繰り返し行った．このような手順で最終的に得られたモデルの，学習・検証用データにおける精度分析結果を表2に示す．なお，表2のデータ数と先の表1のデータ数が一致していないのは，ヨーグルト愛好者のデータをオーバーサンプリングし，その他の顧客のデータ数と同程度にしているためである．

この結果では全体の正分類率が81.8%となっており，かつヨーグルト愛好者の正分類率も86.3%となっていることから，全体的な精度およびヨーグルト愛好者を特定するという目的の双方に関してある程度有望なモデルが構築できたと判断できる．決定木では得られたモデルのルールを参照して解釈を行うことも可能であるが，ここでは事例が架空のものであり，データも限定的なものであるため，得られたルール自体の検討については割愛する．

モデルのパフォーマンス評価には，累積ゲインチャートも利用される．図2は上記モデルを学習・検証用データに適用した結果に基づく累積ゲインチャートである．図の横軸の数値はスコア上位から順にデータを

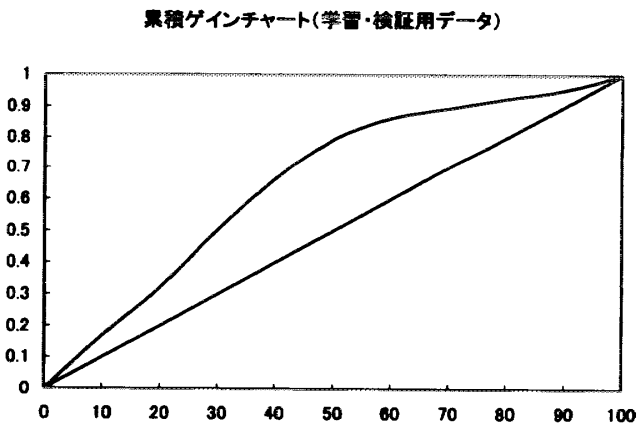


図2 学習・検証用データに基づく累積ゲインチャート

表2 学習・検証用データに基づくコンフュージョン・マトリックス

		予測値		
		ヨーグルト愛好者	その他	計
実績値	ヨーグルト愛好者	309 86.3%	49 13.7%	358 100.0%
	その他	82 22.8%	278 77.2%	360 100.0%
計		391	327	718

全体	正分類	587	81.8%
	誤分類	131	18.2%

並べた場合のランクを示している．つまり10は，スコア上位から10%の顧客データであることを示している．一方，図の縦軸の数値は，ヨーグルト愛好者全体の中での割合を示している．図中の対角線はベースラインであり，これに対して構築したモデルがどれだけよく機能しているかを評価する．例えば図2では，スコア上位40%の顧客をターゲットとした場合，ヨーグルト愛好者全体の60%以上を特定することができ，上位60%の顧客をターゲットとした場合，80%以上のヨーグルト愛好者を特定できていることを示している．累積ゲインチャートでモデルの評価を行う場合，モデルによる曲線とベースラインで囲まれた面積，およびモデルによる曲線の立ち上がりに注目する．モデルに基づく曲線とベースラインで囲まれた部分の面積が大きくなるほど，またモデルに基づく曲線の立ち上がりが急になるほど，当該モデルがよく機能していることを示している．

4.4 モデルのパフォーマンスの評価

構築した決定木モデルを学習・検証用データに適用し，コンフュージョン・マトリックスと累積ゲインチャートで評価した結果は概ね良好なものであった．しかし，当該モデルが学習・検証用データに過度に適合している可能性もある．このような点を考慮し，モデルの学習には用いていない評価用データを使って，最終的なモデルのパフォーマンス評価を行う．

表3は構築した決定木モデルを評価用データに適用した結果から得られたコンフュージョン・マトリックスである．全体の正分類率は90.9%であり，ヨーグルト愛好者の正分類率は100%となっている．また，図3は評価用データにおける累積ゲインチャートである．スコア上位30%のデータにヨーグルト愛好者が100%含まれていることが分かる．これらの結果から，ここで構築されたモデルのパフォーマンスは良好であり，潜在的ヨーグルト愛好者の特定を行うために利用可能であると判断できる．

本事例ではこのモデルを実際のマーケティング・アクションに適用し，その結果を評価することはできないが，仮に特定のマーケティング・アクションに適用する場合，ヨーグルト非購買者の顧客データに，上記の最終的なモデルを適用しスコアを算出した上で，予算制約に基づき許容される人数を，スコア上位顧客から抽出しターゲットとすることとなるであろう．

5. まとめ

本稿では，マーケティング研究におけるマーケットセグメンテーションとターゲティングに関して概観した上で，特にダイレクト・マーケティングあるいはロイヤルティ・マーケティングにおける顧客ターゲティングに焦点を当て，Berry and Linoff[1]が提示している分析プロセスに従い分析事例を交えながら，その解説を行った．分析事例ではモデル構築手法として決定木を例に，潜在的ヨーグルト愛好者の特定を行うためのモデル構築について説明した．得られたモデルを学習には用いていない評価用データに適用し，パフォーマンスの評価を行った結果，全体的な正分類，ヨーグルト愛好者の正分類率，累積ゲインにおいて良好な結果を示した．

これらのデータマイニング手法を利用した顧客タ―

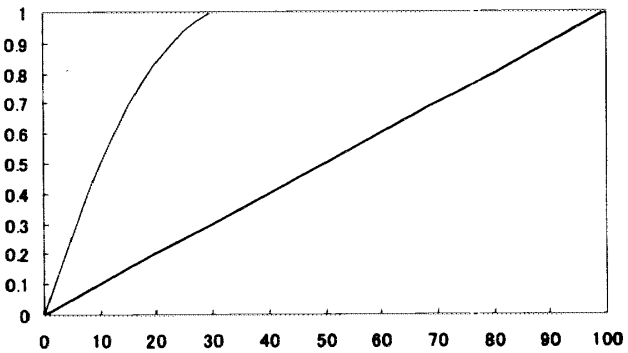


図3 評価用データに基づく累積ゲインチャート

表3 評価用データに基づくコンフュージョン・マトリックス

		予測値		
		ヨーグルト愛好者	その他	計
実績値	ヨーグルト愛好者	6 100.0%	0 0.0%	6 100.0%
	その他	3 11.1%	24 88.9%	27 100.0%
計		9	24	33

全体	正分類	30	90.9%
	誤分類	3	9.1%

ゲティング分析の説明に際しては，可能な限りプロセスを単純化して，内容も比較的平易なものとすることを心掛けたつもりである．しかしそのために，データの準備段階についての説明やデータマイニングの各種手法に関する解説，ニューラルネットなど他の手法の適用事例，複数の予測モデルを構築した上でのモデル選択などの話題については取り上げていない．これらの話題についての詳細に興味がある場合は，参考文献として記載したものなどを参照いただきたい．

最後にこのような分析を実際に行う場合，データマイニングのプロセスを幾度も反復しながら分析作業を行うことによって，最終的に満足できる結果を得ることができるということは十分認識しておいていただければ幸いである．

参考文献

[1] M. J. A. Berry and G. S. Linoff, *Mastering Data Mining, The Art and Science of Customer Relationship Management,* John Wiley & Sons, 2000 (江原・金子・斎藤・佐藤・清水・寺田・守口共訳,『マスタリング・データマイニング CRMのアートとサイエンス 理論編』, 海文堂出版, 2002. ；江原・斎藤・佐藤・清水・守口共訳,『マスタリング・データマイニング CRMのアートとサイエンス 事例編』, 海文堂出版, 2002.).

[2] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques, Sales and Customer Support,* John Wiley & Sons, 1997 (SASインスティチュートジャパン・江原・佐藤共訳,『データマイニング手法 営業，マーケティング，カスタマーサポートのための顧客分析』, 海文堂出版, 1999.).

[3] T. M. Mitchell, *Machine Learning,* McGraw-Hill, 1997.

[4] W. Smith, "Product Differentiation and Market Segmentation As Alternative Marketing Strategies", *Journal of Marketing,* Vol. 21, pp. 3-8, 1956.

[5] M. Wedel and W. A. Kamakura, "The Historical Development of The Marketing Segmentation Concept", *Market Segmentation, Conceptual and Methodological Funndations,* Kluwer Academic Publishers, 1998.

[6] 片平秀貴,『マーケティングサイエンス』, 東京大学出版会, 1987.

[7] J. R. Bult and T. Wansbeek, "Optimal Selection for Direct Mail", *Marketing Science,* Vol. 14, No. 4, pp. 378-394, 1995.

[8] R. E. Rossi, R. E. McCulloch and G. M. Alleby, "The Value of Parchase History Data in Target Marketing", *Marketing Science,* Vol. 15, No. 4, pp. 321-340, 1996.

[9] M. M. Long and L. G. Schiffman, "Consumption Values and Relationships: Segmenting The Market for Frequency Programs", *Journal of Consumer Marketing,* Vol. 17, No. 3, pp. 214-232, 2000.