Department of Social Systems and Management

Discussion Paper Series

No. 1123

CRM のための優良顧客識別手法の特性評価と財務効果

by

鈴木 秀男, 水野 誠, 住田 潮, 佐治 明

June 2005

UNIVERSITY OF TSUKUBA

Tsukuba, Ibaraki 305-8573

JAPAN

CRM のための優良顧客識別手法の特性評価と財務効果

鈴木 秀男　　水野 誠　　住田 潮　　佐治 明
筑波大学

和文概要　　CRM (Customer Relationship Management) において,「優良顧客の識別」はますますその重要性を高めつつある．従来の識別分析においては，識別の正確さを示す正答率を評価の基準とすることが多かった．しかし，識別された優良顧客に対してプロモーションを実施しその財務効果を測ることを考えると，以下の 2 種の過誤を区別することは重要である．すなわち，優良顧客を優良でない顧客として識別してしまう第 1 種の過誤は機会損失の増大をもたらし，優良でない顧客を優良顧客と識別してしまう第 2 種の過誤は非効率的なキャンペーン費用の増大をもたらす．本論文では，これら 2 種の過誤を区別できない正答率による評価の弱点を克服するため，情報検索の分野で広く用いられている 2 つの指標，リコールと精度を導入する．識別された優良顧客に対するプロモーションの財務効果をそれらの関数としてモデル化することにより，様々な識別手法のパラメータ選択並びに比較評価に対し，経営的視点を取り入れた新たな方法論を確立する．実際の POS データを用い，この方法論を単純予測・決定木・ロジスティック回帰分析・SVM の 4 手法に対して適用し，各手法の識別特性を比較分析する．

キーワード: マーケティング，CRM，優良顧客識別，リコール，精度，データ解析

1. はじめに

ここ 10 年間，あらゆる産業分野においてインターネットによるビジネス・モデル革命が急速に進行している．逆オークション方式の導入により遊休資源を活性化させるインターネットに固有の新しい市場形態が生み出される一方，旧来型の企業活動を著しく簡素化し効率化するビジネス・モデルが数多く開発されつつある．インターネットによるこうしたビジネス・モデル革命は，総じて e-ビジネスと呼ばれ，マーケティング分野においても e-マーケティングが主流となりつつある．インターネットの出現以前では，テレビ・新聞・ラジオといったマス・メディアを利用した一方向のマス・マーケティングが大規模に展開される一方，電話による聞き取り調査・出口調査を代表とする双方向の 1 対 1 マーケティングが小規模に行われていた．しかし，双方向の 1 対 1 マーケティングを大規模に展開することは，時間と費用の両面で非常に困難であった．e-マーケティングの最大の特徴は，この困難さを克服し，迅速且つ廉価にマス・マーケティングと 1 対 1 マーケティングを同時に実現することを可能にした点にある．

こうした基本的動向の中で，顧客データベースを活用した CRM (Customer Relationship Management) がますます重要視されてきている．CRM とは，個々の顧客と企業とが，双方向で情報の遣り取りを継続する密接な関係を維持し，顧客の生涯価値を高めるための経営手法である．CRM に付いては，例えば，古林 [8] に詳しい．CRM の主要な機能として，顧客属性や購買データの収集，データベースの構築，データ分析に基づくアクションプランの作成・実施等が挙げられるが，とりわけ重要となるのが，優良顧客の識別分析である．RFM (Recency, Frequency, Monetary) 分析やクロス分析といった顧客の順位付けに関する分析を通して，利益貢献度が高いとされる顧客層を探り当て，将来の購買行動を予測することに

よって優良顧客を識別する．さらに，識別された優良顧客に対してプロモーションを実施する等のアクションを起こし，優良顧客の維持や収益の増加を目指すことになる．

顧客の識別手法としては，データマイニングの分野でよく用いられる決定木やニューラルネットワーク，統計的手法である判別分析やロジスティック回帰分析，あるいは，近年，脚光を浴びつつあるSVM (Support Vector Machine) 等，様々なものが存在する．これに伴い，個々の手法に必要とされるパラメータの値をどのように設定するか，さらには，異なる識別手法をどのように評価し選択するか，という問題が生じる．従来の研究では，識別がどのくらい正しく行われたかを割合で示す正答率が評価の基準とされてきた．しかし，全体の正答率のみによっては,『優良顧客をそうでないと判断してしまう誤り』と『非優良顧客を優良顧客と判断してしまう誤り』を区別できない．CRMにおいては，この2種の誤りを区別することは重要であり，そのためには，情報検索の分野で良く用いられる2つの指標，リコール (Recall) と精度 (Precision) [1], [7] を導入することが有効である．優良顧客識別問題の観点から解釈すれば，リコールは実際の優良顧客をどの程度正しく識別しているかを表す指標であり，精度は優良顧客として識別した顧客の中でどの程度実際の優良顧客が含まれていたかを表す指標ということになる．また，識別された優良顧客に対してプロモーションを行う際には，その利益効果により識別手法の評価を行うことも必要である [3].

本論文の目的は，識別手法の一般的構造に数学的表現を与え，識別された優良顧客に対するプロモーションの財務効果をリコールと精度の関数としてモデル化し，識別手法の比較評価を行う新たな方法論を確立することにある．優良顧客を優良でない顧客として識別してしまう誤りを第1種の過誤，優良でない顧客を優良顧客と識別してしまう誤りを第2種の過誤と定義し，識別された優良顧客に対するプロモーションの財務効果を考えると，第1種の過誤は機会損失の増大をもたらし，第2種の過誤は非効率的なキャンペーン費用の増大をもたらす．優良顧客の識別基準を緩めると，第1種の過誤が小さくなる一方，第2種の過誤が大きくなる．逆に，優良顧客の識別基準を厳しくすると，第1種の過誤が大きくなり，第2種の過誤が小さくなる．このトレード・オフ効果を評価するため，プロモーションの結果に対する利益関数を導入し，その財務効果をモデル化する．実際のPOSデータを用い，この方法論を単純予測・決定木・ロジスティック回帰分析・SVMの4手法に対して適用し，数値的に最適識別基準を求めることによって4手法の識別特性を比較分析する．

本論文の構造は，以下の通りである．第2章では，一般的な優良顧客識別手法を数学的に表現する．コンフュージョン・マトリックス，正答率，リコール，精度といった基礎概念を導入し，正答率を一定にした場合のリコールと精度の関係を構造的に明らかにする．第3章でプロモーションの結果に対する利益関数を導入し，リコールと精度に基づいてプロモーションの財務効果をモデル化する．第4章では実証分析に用いるPOSデータの内容を記述し，分析の対象として有効と認められる説明変数を確定する．これらの説明変数から構成される顧客の属性ベクトルを導入し，単純予測・決定木・ロジスティック回帰分析・SVMの4手法に対して数値実験の構造を決定する．第5章では，これら4手法に関する数値実験の結果を報告する．各手法に対し最適識別基準を求めるとともに，各手法の識別特性を比較分析する．最後に，第6章で結論を纏める．読者の便宜を図るため，付録A, Bで顧客の属性ベクトルを構成する説明変数とその概要を示し，付録Cでは決定木・ロジスティック回帰分析・SVMの概略を纏めておく．

2. 優良顧客識別手法の構造的特性

N 人からなる顧客の集合 $CS = \{c_1, \ldots, c_N\}$ を考える．各顧客 c_i に対し性別・年齢や過去の購買行動等の説明変数からなる属性ベクトルを $\boldsymbol{x}_i$ によって表し，その定義域を Ω とする．すなわち，$\boldsymbol{x}_i \in \Omega, 1 \leq i \leq N$. 各 c_i は，予め定められた基準と次期購買行動の結果によって，優良顧客か非優良顧客かを決定されるものとする．この判別構造を写像 $D^* : C \to \{-1, 1\}$ で表す．優良顧客の集合を G, 非優良顧客の集合を B とすると，

$$CS = B \cup G \tag{2.1}$$

と書けて，

$$B = \{c_i : D^*(c_i) = -1\};\ G = \{c_i : D^*(c_i) = 1\} \tag{2.2}$$

となる．優良顧客と非優良顧客の数をそれぞれ

$$X_B = |B|;\ X_G = |G| \tag{2.3}$$

で表す．ここで，B と G は各顧客の次期購買行動が判明した上で初めて決定されることを改めて注意しておく．従って，現時点における問題は，属性ベクトル $\boldsymbol{x}_i$ に基づいて各顧客の次期購買行動を予測し，B に属するか G に属するかを推定することにある．

上述した推定の仕組を識別関数 $D : \Omega \to [0, 1)$ と識別基準 $z \in [0, 1]$ で表す．すなわち，

$$B(z) = \{c_i : D(\boldsymbol{x}_i) < z\};\ G(z) = \{c_i : D(\boldsymbol{x}_i) \geq z\} \tag{2.4}$$

と定義すると，

$$CS = B(z) \cup G(z) \tag{2.5}$$

が成立する．(2.4) の定義から，$G(z)$ と $B(z)$ は明らかに

$$\begin{cases} G(0) = CS;\ G(1) = \emptyset;\ z_1 < z_2 \Rightarrow G(z_1) \supset G(z_2) \\ B(0) = \emptyset;\ B(1) = CS;\ z_1 < z_2 \Rightarrow B(z_1) \subset B(z_2) \end{cases} \tag{2.6}$$

という性質を有する．写像 D と z は一般的な識別手法の推定構造を表現し，$D(\boldsymbol{x}_i)$ を評価し z と比較することによって，G あるいは B に入ると推定される顧客の集合 $G(z)$ と $B(z)$ が決定されることになる．本論文では，$G(z)$ を z に対するターゲット顧客集合と呼ぶ．

一般的には，ターゲット顧客集合 $G(z)$ と優良顧客集合 G とは必ずしも一致しない．この違いを捉えるため，以下のセル関数を導入する．

$$\begin{aligned} x_{BB}(z) &= |B(z) \cap B|; \quad x_{BG}(z) = |B(z) \cap G|; \\ x_{GB}(z) &= |G(z) \cap B|; \quad x_{GG}(z) = |G(z) \cap G|. \end{aligned} \tag{2.7}$$

これらのセル関数は，データマイニング分野において良く知られるコンフュージョン・マトリックスを構成する（例えば，Berry and Linoff [3] を参照のこと）．コンフュージョン・マトリックスを表 1 に示す．

表 1: コンフュージョン・マトリックス

		実際		
		B	G	計
識別	$B(z)$	$x_{BB}(z)$	$x_{BG}(z)$	$X_{B(z)}$
	$G(z)$	$x_{GB}(z)$	$x_{GG}(z)$	$X_{G(z)}$
計		X_B	X_G	N

ここで,

$$\begin{aligned} x_{BB}(z) + x_{GB}(z) = X_B; \quad & x_{BG}(z) + x_{GG}(z) = X_G; \\ x_{BB}(z) + x_{BG}(z) = X_{B(z)}; \quad & x_{GB}(z) + x_{GG}(z) = X_{G(z)} \end{aligned} \tag{2.8}$$

と定義され，また,

$$X_B + X_G = X_{B(z)} + X_{G(z)} = N \tag{2.9}$$

となることに注意しておく.

(2.6) と (2.7) より，$x_{ij}(z),\ i,j \in \{B,G\}$ は以下の性質を満たす.

$$\begin{gathered} x_{BB}(z) \text{ 及び } x_{BG}(z) \text{ は } z \in [0,1] \text{ に関して非減少であり}, \\ \lim_{z\to 0+} x_{BB}(z) = 0,\ \lim_{z\to 1-} x_{BB}(z) = X_B, \\ \lim_{z\to 0+} x_{BG}(z) = 0,\ \lim_{z\to 1-} x_{BG}(z) = X_G. \end{gathered} \tag{2.10}$$

$$\begin{gathered} x_{GB}(z) \text{ 及び } x_{GG}(z) \text{ は } z \in [0,1] \text{ に関して非増加であり}, \\ \lim_{z\to 0+} x_{GB}(z) = X_B,\ \lim_{z\to 1-} x_{GB}(z) = 0, \\ \lim_{z\to 0+} x_{GG}(z) = X_G,\ \lim_{z\to 1-} x_{GG}(z) = 0. \end{gathered} \tag{2.11}$$

従来，識別手法の性能評価は正答率

$$A(z) = \frac{x_{BB}(z) + x_{GG}(z)}{N} \tag{2.12}$$

に拠ることが一般的であった．しかし，優良顧客を優良でない顧客として識別してしまう誤りを第 1 種の過誤，優良でない顧客を優良顧客と識別してしまう誤りを第 2 種の過誤と定義すると，正答率のみではこれら 2 種の過誤を区別できないことになる．識別された優良顧客に対してプロモーションを行うことを考えると，第 1 種の過誤は機会損失の増大をもたらし，第 2 種の過誤は非効率的なキャンペーン費用の増大をもたらす．従って，識別手法を財務効果の観点から評価するためには，これら 2 種の過誤を峻別することが重要となる．本論文では，情報検索の分野で良く知られている 2 つの評価基準，リコールと精度，を用いることにより，この問題を克服する．リコールと精度は，次式によって定義される.

$$R(z) = \frac{x_{GG}(z)}{X_G};\ P(z) = \frac{x_{GG}(z)}{X_{G(z)}}. \tag{2.13}$$

リコール $R(z)$ は実際の優良顧客をどのくらい正しく識別しているかを表す指標であり，精度 $P(z)$ は優良顧客として識別した顧客の中でどのくらい実際の優良顧客が含まれていたかを表す指標ということになる．(2.13) より，$R(z)$ と $P(z)$ の間には，次の関係が成立する．

$$\frac{R(z)}{P(z)} - \frac{X_{G(z)}}{X_G} = 0. \tag{2.14}$$

(2.11) と (2.13) より，

$$\begin{aligned} &R(z) = \frac{x_{GG}(z)}{X_G} \text{ は } z \in [0,1] \text{ に関して非増加であり，} \\ &R(0) = 1 \text{ 且つ } R(1) = 0 \end{aligned} \tag{2.15}$$

が成立し，

$$\lambda = \frac{X_G}{N} \tag{2.16}$$

と定義すると，

$$P(z) = \frac{x_{GG}(z)}{X_{G(z)}} = \frac{1}{1 + r_G(z)};\ \ r_G(z) = \frac{x_{GB}(z)}{x_{GG}(z)};\ \ P(0) = \lambda \tag{2.17}$$

と書ける．$P(z)$ は，$z = 1$ では定義されない．その極限値 $P(1_-)$ と単調性は，$r_G(z)$ の挙動によって左右される. すなわち，

$$\begin{aligned} &\text{もし } r_G(z) \text{ が } z \in [0,1] \text{ に関して非増加であるならば，} \\ &P(z) \text{ は } z \text{ に関して非減少で且つ } P(1_-) = \frac{1}{1 + r_G(1_-)} \end{aligned} \tag{2.18}$$

が成立する．

正答率 $A(z)$, リコール $R(z)$, 精度 $P(z)$ の間には，以下の関係が成立することを Alvarez [1] が示している．

$$A(z) = \frac{1}{N}\left\{X_B + X_G R(z)\left(2 - \frac{1}{P(z)}\right)\right\}. \tag{2.19}$$

この関係は，(2.8), (2.9), (2.12) と (2.13) より導かれる．この関係を図 1 に示す．正答率を一定にした場合，精度が $\frac{1}{2}$ 以下の範囲では精度はリコールの単調増加関数であり，$\frac{1}{2}$ 以上では単調減少関数となっている点に注意しておく．本論文では，プロモーションの財務効果に対する $R(z)$ と $P(z)$ の果たす役割を分析し，単純予測・決定木・ロジスティック回帰分析・SVM の 4 手法に関して最適識別基準 z^* を数値的に求め，4 手法の識別特性を比較分析することに主眼を置く．

3. マーケティング・キャンペーンの財務効果

識別手法と識別基準 $z \in [0,1]$ が与えられたとき，マーケティング・キャンペーンを $G(z)$ の顧客に対してのみ実施するという戦略を考える．本論文の主な目的は，この戦略の下で識別手法と識別基準 z を変化させるとき，キャンペーンの財務効果がどのような影響を受けるかを分析することにある．

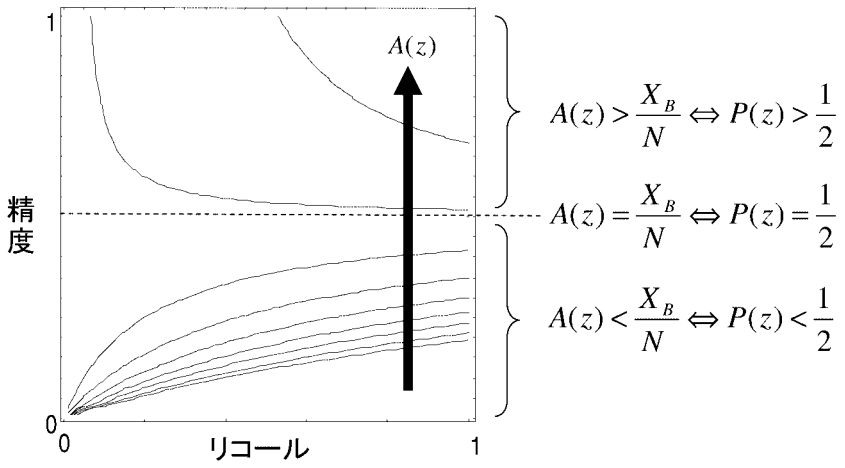


図 1: 正答率等高線

θ_B を非優良顧客一人当たりの期待収入とし，優良顧客と推定された非優良顧客は，プロモーションの影響により θ_B から $\theta_B(1+\eta_B)$ へと支出を増加させると仮定する．ここで，$\eta_B \in [0,\infty)$ としておく．優良顧客に対して θ_G 及び η_G を同様に定義する．明らかに，優良顧客は非優良顧客よりも多くの金額を支出する．また，プロモーションによる支出の増分も優良顧客の方が大きいものとする．従って以下では，

$$\theta_B < \theta_G \quad \text{及び} \quad \theta_B\eta_B < \theta_G\eta_G \tag{3.1}$$

を仮定する．顧客一人当たりのプロモーション費用を $\nu > 0$ で表すと，プロモーション実施時の利益は，

$$\begin{aligned} V(z) = {} & \theta_B x_{BB}(z) + \{\theta_B(1+\eta_B) - \nu\} x_{GB}(z) \\ & + \theta_G x_{BG}(z) + \{\theta_G(1+\eta_G) - \nu\} x_{GG}(z) \end{aligned} \tag{3.2}$$

と表せる．ここで $\eta_B = \eta_G = \nu = 0$ の場合，$V(z)$ はプロモーションを実施しない際の利益となる．従ってプロモーションの財務効果は，それらの差，すなわち，

$$\Delta V(z) = (\theta_B\eta_B - \nu) x_{GB}(z) + (\theta_G\eta_G - \nu) x_{GG}(z) \tag{3.3}$$

によって測定することができる．この $\Delta V(z)$ に基づき，識別基準 z の設定，さらには識別手法間の優劣比較を検討することが可能となる．以下では，$\Delta V(z)$ を z に対する財務効果と呼ぶことにする．

$\Delta V(z)$ は，(2.8), (2.13) と (3.3) により，リコール $R(z)$ と精度 $P(z)$ の関数として以下のように与えられる．

$$\Delta V(z) = X_G R(z) \left(\gamma_B \frac{1 - P(z)}{P(z)} + \gamma_G \right). \tag{3.4}$$

ここで，

$$\gamma_B = \theta_B\eta_B - \nu \quad \text{及び} \quad \gamma_G = \theta_G\eta_G - \nu \tag{3.5}$$

である．なお，(3.1) と (3.5) から，$\gamma_B < \gamma_G$ が成立する．

(3.4) より，財務効果を一定としたときの $R(z)$ と $P(z)$ の関係を財務効果等高線として描くことができ，その一例を図 2 に示す．財務効果は，損益分岐直線 $(\Delta V(z) = 0)$ の上半部では正であり，下半部では負である．財務効果等高線の形状や損益分岐直線の位置は，期待収入 (θ_B, θ_G)，プロモーションへの反応 (η_B, η_G)，プロモーション費用 (ν) に依存する．

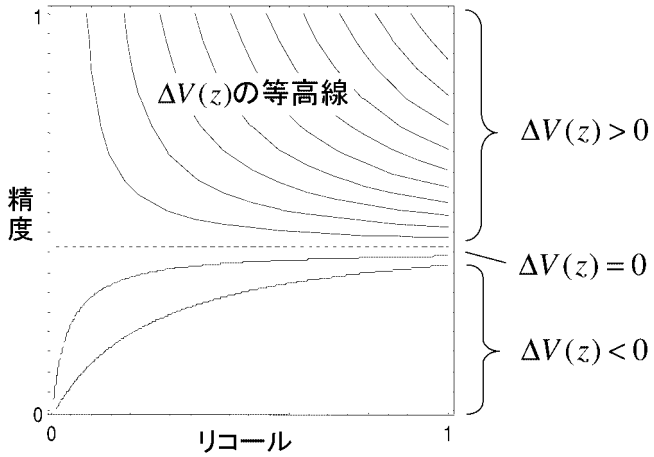


図 2: 財務効果等高線

図 1 で見たように，正答率 $A(z)$ を一定とし $P(z) > \frac{1}{2}$ が成立するとき，$R(z)$ と $P(z)$ の間には反比例関係が成立する．情報検索の分野では，この性質に着目して $R(z)$ と $P(z)$ を複合化した指標 (例えば，F 尺度=$2R(z)P(z)/\{R(z) + P(z)\}$) 等が提案されている [1], [7]．財務効果 $\Delta V(z)$ は，(3.4) により $R(z)$ と $P(z)$ の関数として表現されており，F 尺度と同様に両者を複合化した指標である．しかし，F 尺度が経済的意味を持たないことに比べ，財務効果 $\Delta V(z)$ は経営的視点を取り入れた指標となっている点で著しく異なっている．

或る識別手法が与えられたとき，その財務効果 $\Delta V(z)$ は識別基準 $z \in [0, 1]$ の関数として与えられる．従って，以下の最大化問題を解くことにより，最適識別基準 $z^* \in [0, 1]$ を求めることができる．

最大化問題 3.1

$$\Delta V(z^*) = \max_{0 \le z \le 1} \left[\Delta V(z) = X_G R(z) \left(\gamma_B \frac{1 - P(z)}{P(z)} + \gamma_G \right) \right]$$

を満足する $z^* \in [0, 1]$ を見つけよ．

もし，$0 < \gamma_B < \gamma_G$ であれば，(3.5) より非優良顧客に対する財務効果も正となり，識別せずに全ての顧客に対してプロモーションを行うことが最適となる．この自明の場合を避けるため，本論文では，

$$\gamma_B < 0 < \gamma_G \tag{3.6}$$

を仮定する．

最大化問題 3.1 を解くことは，(2.14) で定まる曲線 $\{(R(z), P(z)) : \ z \in [0, 1)\}$ に対し，$\Delta V(z) > 0$ の範囲における財務効果等高線との接点 z^* を求めることと同値である．その概

念図を図 3 に示す．しかし，この曲線を定める関係式 (2.14) は X_G を含み，顧客の次期購買行動が確定するまで判明しないことから，最大化問題 3.1 を数学的に解くことは不可能である．本論文では，1 期 3ヶ月からなる年間 POS データ I, II, III, IV を用い，データ I に基づいて顧客の属性ベクトルを決定し，データ I・II によって与えられた識別手法に対応する識別モデルを構築する．さらに，データ II・III を用いてコンフュージョン・マトリックスを生成し，最適識別基準 z^* を数値的に求める．最後に，識別モデルと最適識別基準 z^* の財務効果をデータ III・IV によって検証する．詳細に付いては，第 4 章で述べる．

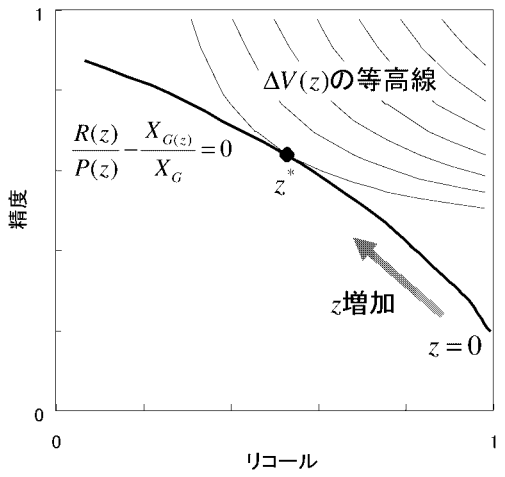


図 3: 最適識別基準 z^* の概念図

4. POS データの概要と 4 つの識別手法

分析対象となるデータは「平成 15 年度データ解析コンペティション」において提供されたドラッグストア顧客 ID 付 POS データであり，顧客の一回の購買内容がカテゴリごとに集計されている．期間は，2002 年 7 月 1 日から 2003 年 6 月 30 日までの 1 年間であり，無作為抽出された顧客 64,453 名 (男性 15.7%，女性 84.3%) を対象とし，96 商品カテゴリ (JICFS 小分類コード)，レコード総量 2,362,163 レコードの規模を持ち，対応する総売上高は 1,555,519,783 円に上る．該当期間を 4 分割し，7-9 月データをデータ I, 10-12 月データをデータ II, 1-3 月データをデータ III, 4-6 月データをデータ IV と呼ぶことにする．

データ I の原データを検討し，先ず，分析目的に有効と思われる説明変数を 36 個作成した．説明変数の作成に当たっては，性別・年齢を始め，直近購買日・購買金額・購買頻度等の RFM 関連項目，商品群の類別化と関連データの集約等に留意した．また，複数の説明変数を組み合わせて一つの指標を作成することも行った．その上で，結果として得られた説明変数の各組み合わせに付いて相関係数を求め，その値が 0.76 以上のものに付いては一方を排除した．最終的には，付録 A に示す 26 の説明変数によって顧客の属性ベクトルを構成することを決定した．次に，各期間に付いて，その期間中に少なくとも 1 回以上購買した顧客を対象として，データ I によって決定された説明変数に基づき，顧客の属性ベクトルを生成する．以後，顧客 c_i の期間 J に対応する属性ベクトルを $\boldsymbol{x}_i(J)$, $J = \mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}$ と書く．もし c_i が期間 J に何も購買しなかった場合には，当該期間の分析から除外するものとする．代表的な説明変数の内容を付録 B で簡潔に纏めておく．なお，以上の分析に当たり，SPSS 社のデータマイニングソフトである Clementine を用いた．

購買金額の高い顧客ほど企業にとって望ましいと考えることは自然であり，優良顧客の集合 G を以下のように定義する．

優良顧客：J 期における優良顧客集合 G とは，その期間中に購買のあった顧客を購買金額の高い順に並べたとき，上位 π%に入る顧客の集合である．

本分析では，紙数の都合から $\pi = 20(\%)$ に付いてのみ検討する．優良顧客集合 G と非優良顧客集合 B に対して，顧客 c_i の判別変数 y_i を次のように定義する．

$$y_i(J) = \begin{cases} 1 & c_i \in G \text{ の場合}, \\ -1 & c_i \in B \text{ の場合}. \end{cases} \tag{4.1}$$

本論文では，識別手法として以下の 4 手法を対象とし，比較検討する．単純予測以外の識別手法の内容に関しては，付録 C に概略を纏めておく．

- 単純予測 (Naïve)
  属性ベクトル $\boldsymbol{x}_i(J)$ の要素である購買金額を用い，相対的位置を基礎に優良顧客の識別を行う．
- 決定木 (C5.0)
  SPSS 社の Clementine 7.1 にある C5.0 を用いて，決定木による識別モデルを構築する．付録 A に示される全ての説明変数を投入し，木のサイズの決定（変数選択）を行う．木のサイズの決定方法は，剪定度=75%（デフォルト），ノードに入る最小サンプル数は文献 [4] を参考にして学習サンプル全体の 0.25%以上と設定する．
- ロジスティック回帰分析 (Logit)
  SPSS 社の Clementine 7.1 にあるロジスティック回帰分析を用いて識別モデルを構築する．付録 A に示される全ての説明変数を投入し，強制投入法を採用する．
- SVM
  Joachims により提供されているソフトウェア SVM^{light}5.00 [14] を用いて線形 SVM モデルを構築する．非線形 SVM に付いては，予備的実験で良い結果が得られなかったので考慮しないこととした．また，目的関数のペナルティ係数 C は，SVM^{light} のデフォルト設定 $C = N/\sum_{i=1}^{N} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i$ に基づいて決定した．説明変数に関しては，上述の決定木で選択されたものとロジスティック回帰分析において影響が大きいとされたものを採用することにした．

数値実験の構造を図 4 に示す．単純予測に付いては，期間 J に対し，$\boldsymbol{x}_i(J)$ の第 k 要素 $x_{ik}(J)$ が購買金額であるとしたとき，$x_{k\cdot\max}(J) = \max_{1\le i\le N}[x_{ik}(J)]$, $x_{k\cdot\min}(J) = \min_{1\le i\le N}[x_{ik}(J)]$ と定義する．このとき，期間 J の識別関数 $D_{\mathrm{I,II}}$ を次式で与える．

$$\text{単純予測:}\quad D_{\mathrm{I,II}}(\boldsymbol{x}_i(J)) = \frac{x_{ik}(J) - x_{k\cdot\min}(\mathrm{I})}{x_{k\cdot\max}(\mathrm{I}) - x_{k\cdot\min}(\mathrm{I})}. \tag{4.2}$$

J が I 以外のとき，$D_{\mathrm{I,II}}(\boldsymbol{x}_i(J))$ は必ずしも $[0,1]$ に収まるとは限らないが，1 を超えた場合には十分小さい $\varepsilon > 0$ に対して $1-\varepsilon$，負となった場合には 0 とする．この調整により，(2.4) と (4.2) から $G(z)$ と $B(z)$ が決定される．

その他の 3 手法に付いては，データ I から属性ベクトル $\boldsymbol{x}_i(\mathrm{I})$ を生成し，データ II によって決定される判別変数 $y_i(\mathrm{II})$ と組み合わせることによって，識別アルゴリズムが決定され，$G(z)$ と $B(z)$ が定まる．決定木に対しては，各 c_i に対してターミナル・ノード s と s におけ

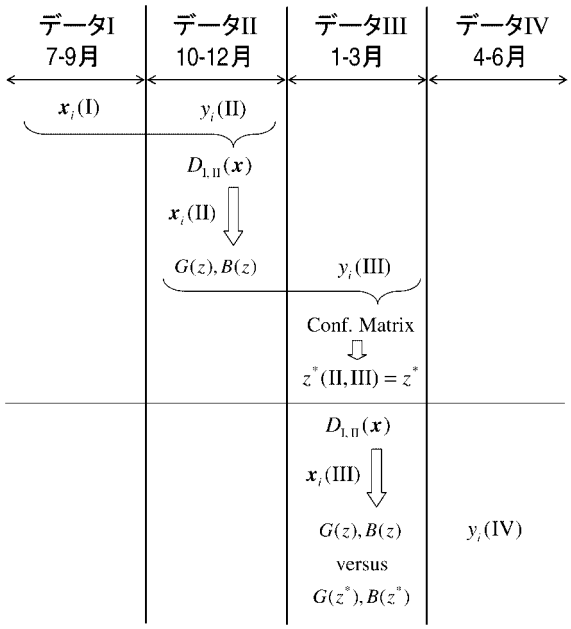


図 4: 数値実験の構造

る確信度 f_s が算出される．このとき，

$$\text{決定木:}\quad D_{\mathrm{I,II}}(\boldsymbol{x}_i(J)) = f_s \tag{4.3}$$

として (2.4) より $G(z)$ と $B(z)$ が確定される．ロジスティック回帰分析では，ベクトル $\boldsymbol{\beta}^*$ と定数 β_0^* が決定され，

$$\text{ロジスティック回帰分析:}\quad D_{\mathrm{I,II}}(\boldsymbol{x}_i(J)) = [1+\exp\{-(\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{x}_i(J)+\beta_0^*)\}]^{-1} \tag{4.4}$$

によって $G(z)$ と $B(z)$ が定まる．SVM では，識別超平面を構成するベクトル $\boldsymbol{w}^*$ と定数 w_0^* が決定され，識別を司る関数 $g(\boldsymbol{x}) = \boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}+w_0^*$ が求まる．ここで，$g(\boldsymbol{x}_{\max}(\mathrm{I})) = \max_{1\le i\le N}[g(\boldsymbol{x}_i(\mathrm{I}))]$，$g(\boldsymbol{x}_{\min}(\mathrm{I})) = \min_{1\le i\le N}[g(\boldsymbol{x}_i(\mathrm{I}))]$ として，

$$\text{SVM:}\quad D_{\mathrm{I,II}}(\boldsymbol{x}_i(J)) = \frac{g(\boldsymbol{x}_i(J)) - g(\boldsymbol{x}_{\min}(\mathrm{I}))}{g(\boldsymbol{x}_{\max}(\mathrm{I})) - g(\boldsymbol{x}_{\min}(\mathrm{I}))} \tag{4.5}$$

を定義する．$\boldsymbol{x}_i$ が期間 I 以外のとき，$D_{\mathrm{I,II}}(\boldsymbol{x}_i)$ は必ずしも $[0,1]$ に収まるとは限らないが，単純予測の場合と同様に，1 を超えた場合には十分小さい $\varepsilon > 0$ に対して $1-\varepsilon$，負となった場合には 0 とする．この調整により，(2.4) と (4.5) から $G(z)$ と $B(z)$ が決定される．

次いで，各手法に付いて，$\boldsymbol{x}_i(\mathrm{I})$ と $y_i(\mathrm{II})$ によって確定された $G(z)$ と $B(z)$ に対し，$y_i(\mathrm{III})$ を用いてコンフュージョン・マトリックスを生成し，z を変化させることによって最適識別基準 z^* を求める．求めた z^* と $D_{\mathrm{I,II}}(\boldsymbol{x}_i(\mathrm{III}))$ によって新たに得られる $G(z^*)$ と $B(z^*)$ を，期間 IV における G と B に対する推定集合として確定し，その財務効果を $y_i(\mathrm{IV})$ を用いて検証し，各手法の比較評価を行う．

5. 数値結果

本章では，単純予測・決定木・ロジスティック回帰分析・SVM の 4 手法に付いて，第 3 章で論じた最適識別基準 z^* を図 4 の手続きに従って推定し，その財務効果を検証する．最

適識別基準の有用性を明らかにするため，識別基準をデフォルト値に設定した場合の財務効果を基準値として採用する．単純予測に対しては $D_{\mathrm{I,II}}(\boldsymbol{x}_i(\mathrm{III}))$ を降順に並べ，上位 $\pi\%$ を識別する z の値をデフォルト値とする．決定木とロジスティック回帰分析では，デフォルト値は $z = 0.5$ と設定される．SVM では，(4.5) に $g(\boldsymbol{x}_i(J)) = 0$ を代入した値，すなわち $z = -g(\boldsymbol{x}_{\min}(\mathrm{I}))/\{g(\boldsymbol{x}_{\max}(\mathrm{I})) - g(\boldsymbol{x}_{\min}(\mathrm{I}))\}$ をデフォルト値とする．

デフォルト値に対する 4 手法の特性を見るため，$y_i(\mathrm{IV})$ によって確定されたリコールと精度を表 2 に示す．単純予測ではリコールと精度がほぼ同じであるのに比べ，他の 3 手法では精度がリコールを大きく上回っている．リコールと精度の双方に付いて同時に他を制する識別手法はなく，リコールと精度の差が最も大きいのが SVM である．

表 2: デフォルト値に基づくリコールと精度

Naïve		C5.0		Logit		SVM	
$R(z)$	$P(z)$	$R(z)$	$P(z)$	$R(z)$	$P(z)$	$R(z)$	$P(z)$
61.37%	61.89%	40.67%	76.30%	36.81%	79.24%	25.50%	85.21%

プロモーションの財務効果を表すパラメータの値に付いては，データ III（1-3 月）の非優良顧客と優良顧客の平均購買金額より θ_B =4,787.5 (円) と θ_G =24,207.8 (円) を求めて III 期の z^* を決定し，最後の検証に際しては，データ IV（4-6 月）に基づき θ_B =4,584.5 (円), θ_G =23,617.7 (円) と設定した．プロモーションは実際には実施されておらず，それが行われた際に顧客の購買額がどれだけ伸びるかは不明である．ここでは，$\eta_B = 0$, $\eta_G = 0.05$ とした．また，X_G は III 期では 8, 292, IV 期では 9, 311 であった．プロモーション費用は，$\nu = 100, 250, 500, 750, 1000$ (円) の 5 つの値を取り上げ，その影響を検証する．

表 3 にプロモーションの財務効果を纏め，表 4 では，最適識別基準に基づいて識別を行った際，$y_i(\mathrm{IV})$ によって確定されるリコールと精度の値を示す．主な結果は以下の通りである．

1) どの識別手法に関しても，最適識別基準に基づくプロモーションの財務効果は，デフォルト値によるそれを大きく上回った
2) 表 3 (c) で示すように，その差をプロモーション費用の関数として見ると，どの識別手法も凸性を持つ
3) プロモーション費用が大きい場合，デフォルト値に基づくプロモーションの財務効果は負となるが，最適識別基準を導入することによって正の値とすることが可能となり，その際にはロジスティック回帰分析もしくは SVM を適用することが望ましい
4) 最適識別基準に基づく識別に関して，ロジスティック回帰分析は決定木を完全に上回る
5) デフォルト値に基づく識別においては，プロモーション費用が 500 円以下である場合には単純予測が最も優れている（これは単純予測のリコールが一番高い理由に拠ると思われる）
6) 最適識別基準に基づく識別を行った際も，プロモーション費用が 250 円以下である場合には単純予測が最も優れており，それが増加するにつれて最適識別手法はロジスティック回帰分析（$\nu \geq 500$）となるが，その差は比較的軽微である
7) 最適識別基準に基づく識別を行った際は，どの識別手法に関しても，プロモーション費用の関数として，リコールは単調減少あるいは非増加であり，精度は単調増加ある

表 3: プロモーションの財務効果

注: 太字は各 ν での最大値を表す.

(a) 最適識別基準に基づく財務効果: $\Delta V(z^*)$ (万円)

ν(円)	Naïve	C5.0	Logit	SVM
100	**757.8**	637.9	755.3	748.1
250	**483.9**	374.7	479.8	472.7
500	217.1	212.2	**221.7**	211.9
750	73.9	75.0	**81.9**	77.7
1000	8.2	0.0	**12.1**	10.5

(b) デフォルト基準に基づく財務効果: $\Delta V(z)$ (万円)

ν(円)	Naïve	C5.0	Logit	SVM
100	**582.4**	397.6	361.4	252.5
250	**443.9**	323.1	296.6	210.7
500	**213.1**	199.1	188.4	141.0
750	−17.7	75.0	**80.3**	71.4
1000	−248.5	−49.1	−27.8	**1.7**

(c) 上記 (a) と (b) の差: $\Delta V(z^*) - \Delta V(z)$ (万円)

ν(円)	Naïve	C5.0	Logit	SVM
100	175.4	240.3	393.9	**495.6**
250	40.0	51.6	183.2	**262.0**
500	4.0	13.1	33.3	**70.9**
750	**91.6**	0.0	1.6	6.3
1000	**256.8**	49.1	39.9	8.8

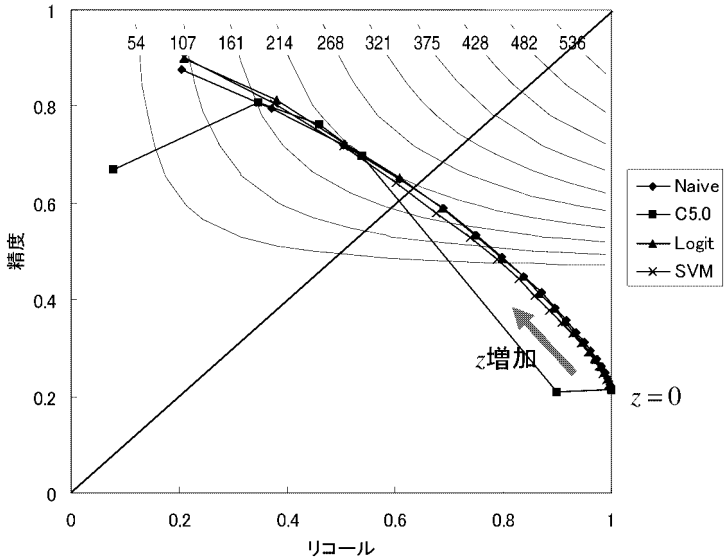
いは非減少である

8) 2) の最小値を達成する ν に対しては，7) では（リコール，精度）の値がデフォルトのそれに近い

図 3 で示した概念図を実際の数値で検証したものを図 5 に示す．すなわち，各識別手法に付いて，リコールと精度の変化を識別基準 z の関数としてグラフ化したものである．これらの曲線が財務効果等高線と接する点が最適識別基準に基づく（リコール，精度）の値に対応する．決定木を除く 3 手法の挙動は概ね似通っており，表 4 で見たように，最適識別基準に基づく（リコール，精度）の値はそれほど変わらない．

表 4: 最適識別基準に基づくリコールと精度

	Naïve		C5.0		Logit		SVM	
ν(円)	$R(z^*)$	$P(z^*)$	$R(z^*)$	$P(z^*)$	$R(z^*)$	$P(z^*)$	$R(z^*)$	$P(z^*)$
100	90.51%	35.51%	100.00%	20.17%	90.02%	35.74%	90.72%	33.87%
250	74.79%	51.44%	48.87%	69.97%	71.94%	53.81%	74.80%	49.78%
500	53.95%	66.78%	48.87%	69.97%	52.99%	68.36%	49.10%	69.70%
750	32.95%	79.79%	40.67%	76.30%	33.11%	81.95%	33.00%	80.83%
1000	12.67%	90.01%	0.00%	–	18.00%	90.20%	17.94%	89.45%



注: 財務効果等高線は，$\eta_B = 0, \eta_G = 0.05, \nu = 500$(円)
の場合で，曲線上の数字はその金額（万円）を表す.

図 5: 識別基準によるリコールと精度の変化

6. 結語

本論文では，識別手法の一般的構造に数学的表現を与え，識別された優良顧客に対するプロモーションの財務効果をリコールと精度の関数としてモデル化し，識別手法の比較評価を行う新たな方法論を提案した．優良顧客を優良でない顧客として識別してしまう誤りを第 1 種の過誤，優良でない顧客を優良顧客と識別してしまう誤りを第 2 種の過誤と定義し，識別された優良顧客に対するプロモーションの財務効果を考えると，第 1 種の過誤は機会損失の増大をもたらし，第 2 種の過誤は非効率的なキャンペーン費用の増大をもたらす．このトレード・オフ効果を評価するため，プロモーションの結果に対する利益関数を導入し，その財務効果をモデル化した．実際の POS データを用い，この方法論を単純予測・決定木・ロジスティック回帰分析・SVM の 4 手法に対して適用し，数値的に最適識別基準を求めることによって 4 手法の識別特性を比較分析した.

デフォルト値に基づいて識別を行う際には，プロモーション費用が低い場合にはリコールが重視され，高いデフォルト・リコール値を持つ単純予測が優越する一方，プロモーション費用が高い場合には精度が重要となり，ロジスティック回帰分析や SVM が優れている結

果となった．最適識別基準に基づくプロモーションの財務効果は，どの識別手法に関しても，デフォルト値によるそれを大きく上回った．この場合，決定木は単純予測，ロジスティック回帰分析，SVM よりほぼ完全に劣っており，一方，3 手法の間には大きな差が見られなかった．

従来の顧客識別においては，2 種の誤りを区別しない正答率のみで評価が行われることが多く，経営的な視点もほとんど含まれていなかった．これに対し，本論文で提案された評価方法は，リコールと精度の関数関係に着目し，1) 2 種の誤りをモデルに陽に取り入れていること，2) プロモーションの財務効果を最終的な判断基準としていること，によって上記の弱点を克服している．

今後の課題としては，顧客クラスの多段階化やプロモーション手段の多様化などを考慮したモデル拡張，また，より豊富な実データに基づく検証の深化等を挙げることができる．

謝辞　データ解析コンペティションを通し，またその後もご親切にアドバイスをして頂きました専修大学の生田目崇氏に心からの謝意を申し上げます．なお，本研究は，文部科学省科学研究費補助金（基礎研究 (C) 17510114）の助成を受けております．

参考文献

[1] S. A. Alvarez: An exact analytical relation among recall, precision and classification accuracy in information retrieval. *Technical Report BC-CS* (Computer Science Department, Boston College, 2002).

[2] 麻生英樹, 津田宏治, 村田昇: パターン認識と学習の統計学 (岩波書店, 2003).

[3] M. J. A. Berry and G. Linoff (著) , 江原淳, 金子武久, 斎藤史朗, 佐藤栄作, 清水聰, 寺田英治, 守口剛 (共訳) : マスタリング・データマイング CRM のアートとサイエンス 理論編 (海文堂, 2002).

[4] M. J. A. Berry and G. Linoff (著) , 江原淳, 斎藤史朗, 佐藤栄作, 清水聰, 守口剛 (共訳): マスタリング・データマイング CRM のアートとサイエンス 事例編 (海文堂, 2002).

[5] M. J. A. Berry and G. Linoff (著) , SAS インスティチュートジャパン, 江原淳, 佐藤栄作 (共訳) : データマイニング手法 (海文堂, 1999).

[6] L. Breiman, J. Friedman, and C. Stone: *Classification and regression trees* (Wadsworth, Inc., 1984).

[7] 岸田和明: 情報検索の理論と技術 (勁草書房, 1998).

[8] 古林宏: CRM の実際 (日本経済新聞社, 2003).

[9] 前田英作: 痛快！サポートベクトルマシンー古くて新しいパターン認識手法ー. 情報処理学会論文誌, **42** (2001), 676-683.

[10] J. N. Morgan and J. A. Sonquist: Problems in the analysis of survey data, and a proposal. *American Statistical Association Journal*, **58** (1963), 415-434.

[11] 中川哲治, 工藤拓, 松本裕治: Support Vector Machine を用いた形態素解析と修正学習法の提案. 情報処理学会論文誌, **44** (2003), 1354-1367.

[12] R. Quinlan: *C4.5: Programs for machine learning* (Morgan Kaufmann, San Mateo, 1993).

[13] 丹後俊郎, 山岡和枝, 高木晴良: ロジスティック回帰分析－ SAS を利用した統計解析の実際－ (朝倉書店, 1996).

[14] T. Joachims: SVM^{light} Support Vector Machine. http://svmlight.joachims.org

付録 A. 顧客の属性ベクトルを構成する説明変数

表 5: 顧客の属性ベクトルの内容

説明変数	データの内容
1	直近購買日
2	2 番目に近い購買日
3	購買回数
4	購買金額
5	RFM ランク
6	1 購買当り購買金額
7	1 購買当り購買個数
8	土日祝日の購買回数
9	土日祝日の購買回数の割合（説明変数 8/説明変数 3）
10	全体との購買類似度
11	売上高トップ 2 カテゴリに対する購買金額
12	売上高トップ 2 カテゴリに対する購買カテゴリ数
13	A カテゴリの購買カテゴリ数
14	B カテゴリの購買金額
15	B カテゴリの購買個数
16	B カテゴリの購買カテゴリ数
17	C カテゴリの購買金額
18	C カテゴリの購買カテゴリ数
19	基礎化粧品デシル
20	メイクアップ化粧品デシル
21	清涼飲料デシル
22	感覚器官及び外皮用薬デシル
23	栄養保健薬デシル
24	衛生紙用品・用具デシル
25	性別
26	年齢

付録 B. 主な説明変数の概要

説明変数 1 (直近購買日) と説明変数 2 (2 番目に近い購買日) は Recency に関するものであり，説明変数 3 (購買回数)，説明変数 8 (土日祝日の購買回数)，説明変数 9 (土日祝日の購買回数の割合) は Frequency, 説明変数 4 (購買金額) と説明変数 6 (1 購買当りの購買金額)

はMonetaryに関わる変数である．なお，説明変数2の「2番目に近い購買日」において，期間中1度しか購買のなかった顧客に付いては，期間最終日と期間初日との期間数を1.1倍した値を与えた．

説明変数5の「RFMランク」とは，以下で定義される3つの優良度（R優良度，F優良度，M優良度）の和を表す．

$$\text{R優良度} = \begin{cases} 1 & (\text{説明変数}1 \leq 5) \\ 0 & (\text{説明変数}1 > 5) \end{cases} \tag{B.1}$$

$$\text{F優良度} = \begin{cases} 1 & (\text{説明変数}3 \geq 7) \\ 0 & (\text{説明変数}3 < 7) \end{cases} \tag{B.2}$$

$$\text{M優良度} = \begin{cases} 1 & (\text{その期の購買金額上位}30\%\text{の顧客}) \\ 0 & (\text{その他}) \end{cases} \tag{B.3}$$

R優良度およびF優良度は，各期間で購買のあった顧客のおよそ30%程度が1を持つように定めている．RFMランクは，RFMという3つの指標から顧客の優良度を評価しており，$0, 1, 2, 3$ のいずれかの値をとる．

説明変数10の「全体との購買類似度」とは，顧客のカテゴリ購買構成比と全体におけるカテゴリ売上構成比との積を各カテゴリに付いて計算し，それらの和をとって100倍した指標である．

説明変数11と説明変数12における「売上高トップ2カテゴリ」とは，各期間における全体の売上高トップ2カテゴリを意味し，説明変数11に関しては各顧客のそのカテゴリに対する購買金額の和，説明変数12に対してはそのカテゴリに対する購買があれば1，なければ0とした上でそれらの和を採ったものである．どの期間においても，基礎化粧品とメイクアップ化粧品が売上高トップ2カテゴリを占めた．

説明変数13から説明変数18の「A,B,Cカテゴリ」に付いては，96あるカテゴリを各期間で売上高の大きい順に並べ，売上構成比が上位75%以内に入るカテゴリをAカテゴリ，75%より大きく90%以内に属するカテゴリをBカテゴリ，それ以外のカテゴリをCカテゴリと定めた．

説明変数19から説明変数24の変数は，6つの特定のカテゴリのデシルである．すなわち，各期間に対象カテゴリを購買した顧客をその購買金額の大きい順に並べた上で10等分し，各区分の代表値として上位から $0.0, \ldots, 0.9$ を割り当て，その顧客が属する区分の代表値をデシル値と定義したものである．期間中に対象カテゴリに属する商品を1度も購買しなかった顧客には，デシル値1.0を与える．選ばれた6つのカテゴリはどの期間でもAカテゴリに属しており，売上構成比の大きいカテゴリである．

識別手法としてSVMを用いる際，本論文ではJoachimsによるソフトウェア SVM^{light} [14] を活用するが，このソフトウェアは，属性ベクトル $\boldsymbol{x}_i$ の定義域である Ω を d 次元立方体に近いものとするよう要望している．各変数の定義域を揃えることによって，外れ値による影響を小さく抑える狙いがあるものと思われる．この要望に応えるため，区間 $[0, 1]$ を定義域として持たない説明変数に関しては，$J = \text{I, II, III, IV}$ に対して以下のスケール変換を行う．

$$J\text{期の}\, c_i\, \text{のスケール変換後の値} = \frac{J\text{期の}\, c_i\, \text{の値} - \text{I期の全顧客の最小値}}{\text{I期の全顧客の最大値} - \text{I期の全顧客の最小値}}.$$

この変換により，各説明変数はI期では0から1の値を持ち，それ以外の期においてもほぼ

0 から 1 の範囲に収まることになる．この変換方法は，明らかに順序関係を保存することに注意しておく．

付録 C.　決定木・ロジスティック回帰分析・SVM の概要

C.1.　決定木

決定木においては，説明変数に基づいて木構造を構成し，属性ベクトルの値によって顧客をルート・ノードから一つのターミナル・ノードへと導くアルゴリズムを決定する．学習データにおける属性ベクトルと判別変数により各ターミナル・ノードでの優良顧客の比率が決定され，その比率をそのノードにおける確信度と定義する．新たな顧客に対しては，属性ベクトルの値に基づきアルゴリズムが作動し，到達するターミナル・ノードが一意的に決定される．そのノードの確信度と識別基準を比較することにより，顧客を類別化することになる．

決定木の長所としては，1) 理解可能なルールに基づくこと，2) 多くの計算を必要とせずに分類を実行できること，3) 連続変数，カテゴリカル変数の双方を扱うことが可能なこと，4) 分岐においてルート・ノードへの近さが影響力の強さを表すと解釈でき，その結果，どの変数が最も重要かを明確に示せること，などが挙げられる [5].

決定木を生成するアルゴリズムとしては様々なものがあり，例えば，CART (Classification and Regression Trees) [6] や CHAID (Chi-squared Automatic Interaction Detector) [10] ，そして C4.5 [12] などがある．

決定木アルゴリズムの基本は，CART に見られるように，データの多様性を最も減少させる分岐を行う点にある．多様性指標が大きければ，各クラスのサンプルが均等であることを意味し，逆にこの指標が小さければ，1 つのクラスにサンプルが集中している状態を表す．多様性指標の例として，エントロピーやジニ係数などがある [5]．データ S をデータ S_{LEFT} とデータ S_{RIGHT} に 2 分割する場合であれば，それぞれの多様性を $M(\cdot)$ とすると，多様性の変化

$$\Delta M = M(S) - \{M(S_{LEFT}) + M(S_{RIGHT})\} \tag{C.1}$$

が最大になるように分岐を繰り返し，木を構成していくことになる．本研究では，決定木として C4.5 アルゴリズムの継承である C5.0 を用いた．

C.2.　ロジスティック回帰分析

ロジスティック回帰分析は，説明変数に基づいて或る関数値を算出し，その値を識別基準と比較することにより判別予測を行う統計的手法である．説明変数としては，連続変数とカテゴリカル変数の双方を扱うことが可能である．このモデルの典型的な利用例としては，或る疫病の生起の有無，地震発生予測などがある [13].

ロジスティック回帰分析では，或る事象（例えば，或る顧客が優良顧客となる事象）の発生する確率 p を d 次元の説明変数 $\boldsymbol{x} = (x_1, ..., x_d)^{\mathrm{T}}$ に基づいて推定する．ここで，説明変数 $\boldsymbol{x}$ の中にカテゴリカル変数（カテゴリー数 $= K$）がある場合は，$K-1$ 個の 2 値変数に変換しておく．説明変数 $\boldsymbol{x}$ の下で事象が発生するという条件付確率 $p(\boldsymbol{x})$ を

$$p(\boldsymbol{x}) = \Pr\{\text{ 発生 } | \, \boldsymbol{x}\} = D(Z) \tag{C.2}$$

と表す．ここで，Z はパラメータ β_0 と $\boldsymbol{\beta} = (\beta_1, ..., \beta_d)^{\mathrm{T}}$ によって

$$Z = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x} + \beta_0 \tag{C.3}$$

と定義され，Z のロジスティック関数 $D(Z)$ は，

$$D(Z) = \frac{\exp(Z)}{1+\exp(Z)} = \frac{1}{1+\exp(-Z)} \tag{C.4}$$

で与えられる．

(C.2)，(C.4) より，

$$\log\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} + \beta_0 \tag{C.5}$$

が成立する．優良顧客の識別に際しては，学習データに含まれる属性ベクトルを用いて (C.5) に基づく回帰分析を行い，パラメータの推定値 $\boldsymbol{\beta}^*$ と β_0^* を求める．属性ベクトル $\boldsymbol{x}$ を持つ新たな顧客に対しては，$p(\boldsymbol{x}) = \boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{x} + \beta_0^*$ を算出し，その値を識別基準と比べることによって類別化を行う．

C.3. SVM

SVM とは，識別問題を解くための学習機械である．以下では主として前田 [9] に従い，クラス間の識別境界線が線形となる線形 SVM の概略を説明する．線形 SVM においては，パラメータ w_0 と $\boldsymbol{w} \in \mathbb{R}^d$ を含む線形関数 $g(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + w_0$ を考える．N 個の属性ベクトル $\boldsymbol{x}_i \in \mathbb{R}^d$ と判別変数 $y_i \in \{-1, 1\}$ $(i = 1, ..., N)$ からなる学習データが与えられたとき，判別変数の値により $\boldsymbol{x}_i$ を 2 つのクラスに識別するようにパラメータを決定することが問題となる．すなわち，全ての i に対して，

$$y_i = f(\boldsymbol{x}_i) = \mathrm{sign}(g(\boldsymbol{x}_i)) \tag{C.6}$$

を満たす $\boldsymbol{w}$ と w_0 を見出すという問題である．勿論，そのようなパラメータが常に存在するとは限らない．

そこで，非負変数 ξ_i $(i = 1, ..., N)$ を用いて，パラメータ $\boldsymbol{w}, w_0$ に関する一意性の制約

$$\forall i,\ g(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + w_0 \begin{cases} \geq 1-\xi_i & (\boldsymbol{x}_i \in \text{クラス } 1) \\ \leq -1+\xi_i & (\boldsymbol{x}_i \in \text{クラス } -1) \end{cases}$$

を与えると，図 6 の超平面 H_1, H_2 間の距離 (マージン) は，$2/\|\boldsymbol{w}\|$ と表せる．SVM では，複数の識別境界候補の中で，マージンを最大にする超平面を最良とみなすので，解くべき最小化問題は，

$$\begin{array}{ll} \underset{\boldsymbol{w},w_0,\boldsymbol{\xi}}{\text{minimize}} & \dfrac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} & \forall i,\ y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + w_0) - (1-\xi_i) \geq 0 \\ & \forall i,\ \xi_i \geq 0 \end{array} \tag{C.7}$$

となる．ここで，目的関数の第 1 項はマージンを大きくとるためのものであり，第 2 項はマージン上及びマージンからはみ出したサンプルに対するペナルティ項である．係数 C は，2 つの項のバランスを決める定数である．

問題 (C.7) から以下の双対問題

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}}{\text{maximize}} & \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} - \dfrac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}D\boldsymbol{\alpha} \\ \text{subject to} & \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{y} = 0 \\ & \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1} \end{array} \tag{C.8}$$

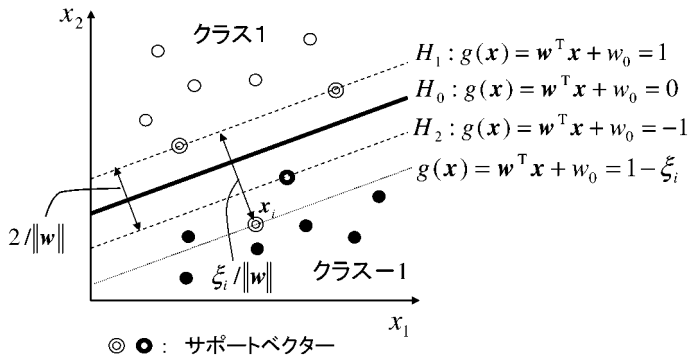


図 6: 線形 SVM

を導出できる．ここで，α_i $(i = 1, ..., N)$ は非負のラグランジュ乗数であり，D はその (i, j) 成分が $y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$ である (N, N) 行列，$\mathbf{1}$ はその全ての成分に 1 を持つ列ベクトルを表す．主問題 (C.7) の最適解を $\boldsymbol{w}^*, w_0^*$, 双対問題 (C.8) の最適解を $\boldsymbol{\alpha}^*$ とすると，線形 SVM における識別関数 $f(\boldsymbol{x})$ は，

$$f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x} + w_0^*) \tag{C.9}$$

$$= \mathrm{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i (\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}) + w_0^*\right) \tag{C.10}$$

となる．ここで w_0^* は，$0 < \alpha_k^* < C$ となる任意の α_k^* に対応する $\boldsymbol{x}_k$ を用いて，

$$w_0^* = y_k - \left(\sum_{i=1}^{N} \alpha_i^* y_i \boldsymbol{x}_i\right)^{\mathrm{T}} \boldsymbol{x}_k \tag{C.11}$$

と求まる．なお，$\alpha_i^* > 0$ に対応する $\boldsymbol{x}_i$ をサポートベクターと呼び，(C.10) と (C.11) より，サポートベクターとなる $\boldsymbol{x}_i$ のみによって，識別関数値が決まることが確認できる．学習データによって $\boldsymbol{\alpha}^*$ を決定し，属性ベクトル $\boldsymbol{x}$ を持つ新たな顧客に対しては，(C.10) によって $f(\boldsymbol{x})$ を算出し，その値が 1 であればクラス 1，-1 であればクラス -1 と判別することになる．

SVM の特徴として，1) マージン最大化基準を採用した識別手法であること，2) 凸 2 次計画問題を解くことにより，最適な識別関数が一意に定まること，3) 線形構造を持たない場合には，属性ベクトルを高次元空間に非線形写像し，写像先の空間において線形識別を行うことにより，元の入力空間での非線形識別を容易に行うことができること，などが挙げられる．SVM は，汎化能力が高く複雑な非線形の識別境界を表現できるにも拘らず，ニューラルネットワークのように局所解問題を心配する必要がないという利点によって，多くの応用研究に活用されている（例えば，[2], [11]）．