# マーケティング情報処理における Kernel-based 手法と GP 手法

時永　祥三

インターネット・マーケティングにおいては，極めて多量の顧客や商品に関する情報を処理する必要がある．いわゆる，データマイニングの応用である．本報告では，特に判別分析とこれに関連する新しい手法を紹介する．マーケティングにおいては，顧客や商品をグループ化することにより，とるべき方策が決まることが多い．顧客情報を入力として，顧客がどのようなクラスに属するかを推定する判別問題に対して，Kernel-based 非線形写像を用いた方法を述べる．また，顧客のクラスを取り出したときに，そのクラスの特徴は何かを求める問題に遺伝的手法（Genetic Programming）を応用する．

キーワード：Kernel-based 手法，判別分析，遺伝的プログラミング，クラスの特徴抽出

## 1. マーケティングと情報処理

近年，コンピュータの能力の飛躍的な向上とインターネット技術の進展にともない，商品販売やこれに関連する顧客の情報を収集することが容易になっている．これにともなって，多量に収集されたデータから，販売に有利な情報を抽出したり，規則性を発見する，いわゆるデータマイニングの方法論が展開されている．データマイニング手法にはさまざまなものが存在し，実際に適用され成果をあげている事例も報告されている．これらをサーベイすることも意義があると思われるが，本稿では，比較的新しい手法とその応用について述べることにする．

従来の販売チャネルに依存する形態と異なり，インターネット・マーケティングにおいては，極めて多数の不特定な顧客を相手にするケースが増大することになる．このようなケースにおいては，顧客をグループとして判別することと，顧客の特徴を抽出することが，販売を促進する上で有効な手段となるであろう．もちろん顧客に限らず，店舗や商品のグループ化特徴抽出への適用も有効であろう．このようなことを考慮して，本稿では判別分析とクラスタ分析に関する新しい手法について述べる．具体的には Kernel-based 手法による非線形判別分析の拡張であり，遺伝的プログラミング（Genetic Programming: GP）手法による言語的な出力を考慮したクラスタ分析，およびクラスタ特徴記述である．

ときなが　しょうぞう
九州大学 大学院経済学研究院経済工学部門
〒812-8581 福岡市東区箱崎 6-19-1

## 2. Kernel-based 手法の原理

### 2.1 データ収集と判別分析

多変量解析の1つの手法として判別分析は多くの分野で用いられており，現在でも金融業においては，企業への貸付や消費者ローンの審査において大きな位置を占めている．判別分析の基本は，観測されたサンプルごとに判別の基礎となるデータを入力変数 $x_i$（判別変数）として与えておき，同時に，このサンプルが所属する分類（これをクラスと呼んでおく）が与えられている場合に，クラスを推定するための $x_i$ に関する線形の関数を求めることである．このように準備されるデータを学習データと呼び，複数のクラスについて同時にデータを収集するので，ペアサンプルとも呼ばれる．

線形の判別関数を求める方法は，クラス内の分散とクラス間の分散の比を最小化するように，パラメータである線形判別関数の係数を求める問題に帰着される．しかしながら線形判別関数によるクラス分けの方法では，関数が線形であることに制約があるため，判別関数を非線形にまで拡張することで，より精度の高い判別を可能とする方法が提案されている．ニューラルネットワークは，このような問題を解決する方法の1つである．ただし，ニューラルネットワークを含めて，非線形判別関数を求める方法では，関数推定の方法論

も非線形問題となるため，計算時間が多大となる問題がある．

このような課題を解決する1つの方法として提案された，代数的に非線形判別関数を求める方法がKernel-based手法である[1]－[4]．Kernel-based手法の特徴は，写像により判別変数の次元を増加させ，より精度の高い判別関数を構成することと，この次元の増加による計算量が抑制できることにある．

### 2.2 Kernel-based手法による判別関数の推定

Kernel-based手法とは，入力変数に対する非線形変換関数を適用し，やや高次元の線形判別関数を用いた問題へと帰着させる方法である．この計算の過程で，dot productという表現を用いる．2次元ベクトル $(x_1, x_2)$ を3次元 $(x_1^2, \sqrt{2}\,x_1x_2, x_2^2)$ という2次モーメントへ変換する変換がある．このとき，2つのベクトル $x=(x_1^2, \sqrt{2}\,x_1x_2, x_2^2)$, $y=(y_1^2, \sqrt{2}\,y_1y_2, y_2^2)$ の要素の，すべての組み合わせの積和 $\sum_{i1,\,i2,\,j1,\,j2=1}^{2} x_{i1}x_{i2}y_{j1}y_{j2}$ を計算することをdot productとよぶ．これを $(x\cdot y)^2=k(x, y)$ として表現する．また，$k=\exp(-\|x-y\|/\sigma)$ の形の計算をする場合にはガウシアンKernelとよぶ．

いま，判別すべきクラスの数を $N$ とし，それぞれのクラス $l$ には $n_l$ 個のサンプルがあり，この合計を $M=\sum_{l=1}^{N} n_l$ とする．$j$ 番目のサンプルを表す変数ベクトルを $x_j$ とする．次のような共分散行列を定義する．

$$C=M^{-1}\sum_{j=1}^{M} x_j x_j^T \tag{1}$$

いま，変換関数 $\phi(x)$ を用いて入力変数 $x$ を変換する．この場合，変換されたデータ $\phi(x)$ の次元 $n$ は，もとの変数の次元 $m$ より多く設定されており（いわゆる高次元），これにより，より自由度の高い判別関数の構成が可能となる．

$$\phi:\ R^m \to F,\ x \to \phi(x) \tag{2}$$

変換されたデータについての共分散行列は，次のようになる．

$$V=M^{-1}\sum_{j=1}^{M} \phi(x_j)\cdot\phi(x_j)^T \tag{3}$$

クラス $l$ のサンプルの入力変数ベクトルの第 $k$ 番目の要素を $x_{lk}$ としておく．この平均値 $\bar{\phi}_l$ を次のように定義し，この平均値に関する共分散行列 $B$ を定義する．

$$\bar{\phi}_l=n_l^{-1}\sum_{k=1}^{n_l} \phi(x_{lk}) \tag{4}$$

$$B=M^{-1}\sum_{l=1}^{N} n_l \bar{\phi}_l \bar{\phi}_l^T \tag{5}$$

変換されたデータについての共分散行列 $V$ は，また次のように書くこともできる．

$$V=M^{-1}\sum_{l=1}^{N}\sum_{k=1}^{n_l} \phi(x_{lk})\cdot\phi(x_{lk})^T \tag{6}$$

線形判別関数を構成することにより，いくつかのクラスに分類する問題は，いわゆるKernel Fisher Discriminant Analysis（KFD）と呼ばれるものであり，複数のクラスの平均値ができるだけ乖離しており，それぞれのクラスの分散は小さくなるように構成する．この問題は，次に示すRayleight coefficientを最大化することに帰着される[1]．

$$\lambda=\frac{v^TBv}{v^TVv} \tag{7}$$

$$v=\sum_{f=1}^{N}\sum_{g=1}^{n_f} \alpha_{fg}\phi(x_{fg}) \tag{8}$$

ここで固有ベクトルは，空間 $F$ の要素の線形結合となることを用いている．$f$ はクラスを指す添え字であり，$g$ はクラス内のサンプルを示す添え字である．

2つのクラス $p, q$ に関して，次のようなdot productを定義する．

$$(k_{ij})_{pq}=\phi(x_{pi})^T\cdot\phi(x_{qi}) \tag{9}$$

これらを用いて，次の $M\times M$ 行列 $K$ を定義する．

$$K=(K_{pq}),\ p, q=1\sim N,$$

$$K_{pq}=(k_{ij})_{pq},\ i=1\sim n_p,\ j=1\sim n_q \tag{10}$$

$K_{pq}$ は $n_p\times n_q$ 行列である．また，クラス $l$ ごとに要素が $1/n_l$ に等しい部分行列 $W_l$ からなる $M\times M$ ブロック対角行列 $W=(W_l),\ l=1\sim N$ を導入する．これにより，式(7)の問題は次のようになる．

$$\lambda=\frac{\alpha^TKWK\alpha}{\alpha^TKK\alpha} \tag{11}$$

$$v^Tv=\alpha^TK\alpha=1 \tag{12}$$

さらに，行列 $K$ を $K=U\Gamma U^T$ のように対角化し，この対角行列 $\Gamma$ を用いてベクトル $\alpha$ の変換を行う．

$$\beta=\Gamma U^T\alpha \tag{13}$$

以上の準備のもとで，KFDアルゴリズムは，次のように整理される．

1．行列 $K, W$ を計算する．

2．固有値分解により行列 $K$ を分解する．

3．固有ベクトル $\beta$ と固有値を計算する．

4．係数 $\alpha$ を用いて固有ベクトル $v$ を計算する．

5．次の式を用いて未知の入力データである変数ベクトル $z$ を有するサンプルが，どちらのクラスタに属するかを，次の関数により推定する．

$$(v^T\cdot\phi(z))=\sum_{f=1}^{N}\sum_{g=1}^{n_f} \alpha_{fg}k(x_{fg}, z) \tag{14}$$

### 2.3 判別分析の応用例

以下では，文献[1]で述べられている例を示すにとどめる．次のような2次元平面に配置される2つのクラスのデータを，それぞれについて200個ずつ人工的に生成する．

Class 1: $X\sim N(0,\sqrt{2})$, $Y_i=X_i^TX_i+N(0,0.01)$

Class 2: $X\sim N(2,0.001)$, $Y\sim(2,0.001)$

この2つのクラスに対してKernel-based手法による判別分析を行う．最初の手法として，$k(x,y)=(x\cdot y)^d$, $d=2$となるKernel functionを用いる．学習データとして20サンプルを用い，判別の検証は200サンプルすべてについて行う．この結果，2つのサンプルを除いて，すべてのサンプルが正確に判別されている．特に，Class 2についてはほとんど1つの点に変換されている．次に，ガウシアンKernelを用いて判別を実施すると，すべてのサンプルが正確に判別されている．

## 3. GPによるクラスタ特徴記述

### 3.1 GPによるルール生成への遺伝的操作

GP手法はこれまで，カオス時系列の予測問題，エージェントによる人工株式市場の分析などへと応用され，有効性が示されている[5]～[9]．GP手法の原理は，多数の関数の木構造（これを個体とよぶ）を初期値として準備しておいて，近似度の高い個体の対に対して交差処理をほどこして，より近似能力の優れた個体を生成することにある．例えば，次の方程式で記述されるロジスティック写像がある．

$$x(t+1)=3.87x(t)[1-x(t)] \tag{15}$$

この変数$x(t)$の初期値に，適当な値を与えて時系列を生成すると，乱雑な時系列となる．あらかじめこの関数により時系列を生成しておいて，このデータを観測値としてGP手法によりもとの関数を推定すると，極めて小さな誤差で達成することができる．

GPのアルゴリズムについては，簡潔に次のように整理される．2つの個体を適合度に比例する確率で選択し，個体が関数を表現する状態を保持するように2つの個体を交差する位置を求め，この交差点の前後を入れ換える．このような交差処理を規定回数繰り返す．このあと，個体のプールの中に存在する相対的に適合度の低い個体を生成された子と入れ換える．以上の処理を十分な回数繰り返す[5]～[9]．図1は交差処理の概要を示す．なお，木構造は等価な前置表現に変換されている．

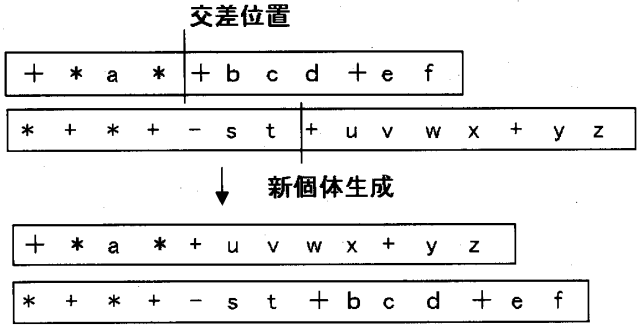


図1 GPにおける交差処理の概要

一方，サンプルがカテゴリ変数で記述されている場合には，判別分析においてもサンプルを特徴付けるプロダクションルール，すなわち論理式を推定することが有効である．この場合，簡単な拡張により，関数近似に用いたGP手法を論理式の推定に応用することが可能である．具体的には，次のような置き換えを行う．

数値型入力変数$v_i$→論理変数$X_i$

算術演算子+, ×→論理演算子And, Or

いま，カテゴリ変数として変数$v_1, v_2, \cdots, v_m$をとり，これらの変数には，それぞれカテゴリ値$s_1, s_2, \cdots$が代入されるとしておく．例えば，カテゴリ変数$v_1, v_2, \cdots, v$がそれぞれ，値$s_3, s_5$をとる場合には次のようになる．

$$v_1=s_3,\ v_2=s_5 \tag{16}$$

したがって，これらを論理命題として結合したものを，プロダクションルールとして記述することができる．さらに簡単化を行うと，論理式は次のような論理変数を含み，これらを論理演算子で結合（GPにおける前置表現ではすべて2項演算に分解されている）した論理式に書き換えられる．

$$X_{kj}=\begin{cases}True, & if \quad v_k=s_j\\ False, & otherwise\end{cases} \tag{17}$$

上に示したようなプロダクションルールの推定方法を，いくつかの代表的な事例に適用し，従来の手法より性能が優れていることが示されている．ただし，これらの事例では，複数のグループを判別する問題に限定されている制限がある[6][7]．

### 3.2 GP手法によるクラスタ特徴記述

クラスタ分析の分野は，大別して，学習データをもとにしてクラスタの代表値などを求め，所属が未知であるサンプルの所属推定をするクラスタ分類と，クラスタとして分離された集合の特徴を分析する方法（以下では，この分野をクラスタ特徴記述とよぶ）とがある．従来の代表的なクラスタ分析手法である多変量解

析法やID3などでは，ペアサンプルとして定義される複数のクラスタ（外的基準として，一方が合格なら他方が不合格であるなどの，区分化されたクラスタ）が必要である点である．したがって，サンプル集合として単独で与えられたクラスタの特徴を抽出する場合には，直接的に適用できない．そのため，外的基準をともなうペアサンプルを必要としないで，かつ，言語的にクラスタ特徴記述が可能な方法が必要となる．

まず，数値的な手法などを用いてデータ全体から特定のクラスタを取り出す．次に，カテゴリカルデータに対する論理変数を仮定し，これら論理変数による論理式をクラスタ特徴記述のルールとしてとらえ，クラスタ内のサンプルだけにヒットする検索ルールへとGP手法を用いて改善する．論理式はGP手法における個体として表現され，プールを構成する．しかしながら，通常のGP手法とは異なり，個体の適合度をクラスタ内部のサンプルへのヒット数に比例するだけではなく，クラスタ以外へのヒット数に反比例するような定義へと変更する．このように適合度の定義を拡張することにより，クラスタ特徴記述を与える論理式を，確実に個体として改善することができる．

いま，GPにおける個体$k$について，データ全体のすべてのサンプルにこの個体で記述される論理式をあてはめ，その論理値が真となる割合によりヒット率を定義する．ただし，注目するクラスタ$c$のほかに，これ以外のクラスタについても調べる必要があるので，クラスタ内$c$でのヒット率と同時に，データ全体でのヒット率を導入している．

ここで，次のような指標を定義する．

$$y_k = T - h_k^2/n_k \tag{18}$$

式(18)に含まれる変数は，以下のように定義される．

$n_k$：全部のサンプルで個体$k$の論理式が真となる数

$h_k$：クラスタ$c$に含まれるサンプルで個体$k$の論理式が真となる数

$T$：クラスタ$c$のサンプル数

個体$k$の適合度$f_k$は，式(18)に示す$y$に正の定数$a$を加えた数の，逆数により定義する．

$$f_k = (a + y_k)^{-1} \tag{19}$$

式(18)に示す指標は，クラスタを特徴付ける論理式がクラスタのサンプルをカバーする割合が大きいほど，ゼロに近くなる．この式(18)の第2項の分母には$n_k$が含まれているが，これは検索のルールが，可能な限りクラスタ$c$内部のサンプルだけをカバーするように調整するためのものであり，クラスタ外のサンプルについても論理式が成り立っている場合には，個体の適合度は低下するようになっている．このようにして個体の適合度が計算されるので，通常のGP手法におけるものと同様に，遺伝的操作を適用し，個体のクラスタ検出能力を改善する．

クラスタ検索のための個体の適合度の最大値が，もはや改善されないことが確認できた時点で，GPによる遺伝的操作を中止する．適合度の最高値が増加しない場合には，適合度が最高となる個体$k$により特徴記述されるレコードの集団，すなわちクラスタが検出・推定されたことに対応している．

### 3.3 German Creditを用いたクラスタ検索

次に，やや実際的なデータに対するクラスタ特徴記述の例をとりあげ，GP手法の適用可能性を議論する．データはドイツの消費者ローン会社で実施された1,000名を対象にした貸付審査の結果データであり，貸付を拒否された300名のデータと，貸付された700名のデータからなる．データの項目は，7つの数値データと，13個のカテゴリカルデータとからなる[9][10]．

このデータの本来の目的は，貸付審査の可否を決めるルールを求めることであるが，シミュレーションではクラスタを分離して，その特徴を記述することに用いる．そのため，最初に全部で1,000個からなるデータからランダムに90個を選択し，次に示す7個の数値型変数を用いて統計パッケージによるクラスタ分析を用いて3つのクラスタを抽出する．

y1：クレジット期間
y2：クレジット額
y3：クレジット利率
y4：現住所での居住期間
y5：年齢
y6：当会社銀行でのクレジット開設数
y7：扶養家族数

この3つのクラスタのそれぞれについて，抽出すべきクラスタ$c$であると仮定し，このクラスタに含まれないサンプルを，クラスタ$c$以外のクラスタ$d$に属するとする．クラスタを抽出するためのカテゴリ変数は，以下のようになる（カッコ内はカテゴリ数）．

x1：手形口座開設の内容(4)
x2：契約継続月数(2)
x3：クレジット履歴(5)
x4：借入目的(11)

表1　$h, n, y, N_{GP}$ の間の関係例

| $N_{GP}$ | 1 | 100 | 200 | 300 | 400 | 500 | 600 |
|---|---|---|---|---|---|---|---|
| $h$ | 6 | 12 | 17 | 23 | 25 | 26 | 30 |
| $n$ | 21 | 14 | 51 | 54 | 37 | 30 | 30 |
| $y$ | 28.2 | 19.5 | 24.3 | 19.9 | 12.7 | 7.7 | 0 |

表2　得られるクラスタ特徴記述の論理式の例

| | | | | | | |
|---|---|---|---|---|---|---|
| And | $x_{53}$ | And | And | $x_{23}$ | $x_{42}$ | $x_{14}$ |
| And | And | $x_{21}$ | $x_{62}$ | Or | $x_{13}$ | $x_{41}$ |
| And | Or | $x_{53}$ | $x_{61}$ | And | $x_{23}$ | $x_{41}$ |
| Or | $x_{62}$ | And | $x_{41}$ | And | $x_{33}$ | $x_{22}$ |

x 5：預金口座内容(5)

x 6：保証人の有無(3)

シミュレーションのための条件は，以下のようにしておく．

個体記述の配列の最大サイズ：$M_s=10$

個体の数プールの大きさ：1,000

表1には，3つのクラスタの1つをクラスタ $c$ とした場合に得られる式(1)に示す $h, n, y$ が最適となる個体の値を，主要なGP世代（$N_{GP}$ ごとに示している）．この表より分かるように，ほぼ第600世代で目的とするクラスタ検索のルールが得られる．その後もGP処理を続けることにより，複数のクラスタ検索ルールが求められる．適合度が最大になる個体によりクラスタ検索が，前置表現のままで得られる．この例を，表2に示している．この表2より分かるように，クラスタの特徴記述として，十分に簡潔な形となっていると言える．

## 参考文献


[1] G. Baudat and E. Anouar, "Generalized disciminant analysis using a Kernel approach," Neural Computing, vol. 12, pp. 2385-2404, 2000.

[2] B. Scholkopf and A. Smola, "Nonlinear component analysis as a kernel eigenvalue problem," Neural Computation, vol. 10, pp. 1299-1319, 1998.

[3] H. Kwon and N. M. Nasrabadi, "Kernel matched subspace detectors for hyperspectral targetdetection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 2, pp. 178-194, 2006.

[4] K. Muller, S. Mika, G. Ratsch, K. Tsuda and B. Scholkopf,, "An introduction to kernel-based learning algorithms," IEEE Trans. on Neural Netwroks, vol. 12, no. 2, pp. 181-201, 2001.

[5] X. Chen and S. Tokinaga, "Multi-agent-based modeling of artificial stock markets by using the co-evolutionary GP approach," JORSJ, vol. 47, no. 3, pp. 163-181, 2004.

[6] 呂　建軍, 時永祥三, "遺伝的プログラミングによる時系列モデルの集合的近似とクラスタリングへの応用," 電子情報通信学会論文誌, vol. J 88-A, no. 7, pp. 803-813, 2005.

[7] 呂　建軍, 時永祥三, "遺伝的プログラミングによる時系列セグメント識別を用いたカテゴリ記号表現に基づく2階層認識手法とその予測への応用," 電子情報通信学会論文誌, vol. J 88-A, no. 11, pp. 1258-1271, 2005.

[8] J. Lu, S. Tokinaga and Y. Ikeda, "Explanatory rule extraction based on the trained neural network and the genetic programming," JORSJ, vol. 43, no. 2, 2006.

[9] 呂　建軍, 時永祥三, "遺伝的プログラミングによるルール生成を用いたクラスタ特徴記述システムの構成とその応用," 電子情報通信学会論文誌, vol. J 89-A, 2006 掲載予定.

[10] http://www.liacc.up.pt/ML/statlog/datasets/german/german.doc.html