DEWS2007 D2-6

カーネル法による現象データマイニングの試み

都築　学† 新美 礼彦† 小西 修†

† 公立はこだて未来大学システム情報科学部 〒 041–8655 北海道函館市亀田中野町 116–2
E-mail: †{c1103075,niimi,okonishi}@fun.ac.jp

あらまし 顧客の識別分析は，顧客に合ったプロモーションを行うことで収益の増加を目的としたものであるが，現在はインターネットのビジネスが普及し，リアルタイムで顧客を識別し，また識別情報も更新する必要が出てきた．従来，顧客を識別するためには，識別に重要な顧客の行動パターンを「ルール」として，顧客の行動パターンの変化の度合いを「スコア値」として用いられてきた．しかし，どのような顧客の行動パターンが顧客の識別に重要であるかは，対象世界の背景知識に頼る部分が大きい．そのため，本論文では顧客の識別に重要な行動パターン (ルール) を，カーネル法により識別されたクラス内から自動的に抽出し，その「ルール」と，振舞いデータから導出した「スコア値」を組み合わせて識別を行うダイナミックなシステムを，現象データマイニングシステムとして提案する．
キーワード 現象データマイニング，カーネル法，振舞いデータ

A phenomenal data mining system based on kernel method

Manabu TSUZUKI†, Ayahiko NIIMI†, and Osamu KONISHI†

† Future University-HAKODATE　Kamedanakano 116–2, Hakodate-shi, Hokkaido, 041–8655 Japan
E-mail: †{c1103075,niimi,okonishi}@fun.ac.jp

Abstract Phenomenal data mining finds relations between the data and the phenomena that give rise to data rather than just relations among the data. For example, suppose POS data does not know customer's behavior. Customer's behaviors are characterized by money, goods, time, tastes, and so on. We define them as behavior data. In this example, the POS data are the data, and the customer's behaviors are phenomena not directly represented in the data. We work mainly with the POS example, but the idea is general. In order to infer phenomena from data, facts about their relations must be supplied. The result of phenomenal data mining might include an extended database with additional fields on existing relations and new relations. Thus the relations describing POS data might be extended with a customer's behavior field, and new relations about important customer's action pattern extracted from each class classified by kernel method. .
Key words phenomenal data mining, behavior data, kernel method

1. はじめに

近年のデータ収集技術の大幅な進歩から，様々な分野で大量のデータが収集・蓄積されるようになった．大量の情報から有用な情報を抽出するデータマイニングは，大量のデータが存在する全ての分野で明らかに有益である．特に，ビジネスの分野では情報こそ競争優位の源泉といわれ，多くの企業はあらゆる情報を蓄積し，これをいかに活用して経営効果・競争優位に結びつけるかが課題となっており，特に，顧客の識別分析が重要視されている．顧客の識別分析は，顧客に合ったプロモーションを行うことで収益の増加を目的としたものであるが，現在はインターネットの普及により，顧客は Web を通して商品を購入したりサービスを受けたりすることが出来る．このような状況では，顧客がホームページに接続すると同時にその顧客を識別し，それぞれの顧客に合った商品やサービスの提示を行うことが必要である．つまり，リアルタイムで顧客を識別し，また識別情報も更新する必要が出てきた．

ここで，識別に重要な情報はこれまでの顧客の行動パターンと，その変化である．例えば,「先月，先々月共に 2〜3 万円購入している」顧客と「先月，先々月共に 10 万円以上購入している」顧客に対して，同様の商品やサービスを提示するのではなく，顧客に合った提示が望ましい．また，顧客の行動パターンの変化は，顧客の購買意欲の増加・低下，異常行動，あるいは興味関心分野の移行を示しており，これに対しても，これまでと同様の商品やサービスを提示するのではなく，顧客に合った提示が望ましい．従来，このような顧客の行動パターンは，顧

客を識別する「ルール」として，顧客の行動パターンの変化は，その変化の度合いをスコアリングした「スコア値」として顧客の識別に用いられてきた．しかし，どのような顧客の行動パターン (ルール) が顧客の識別に重要であるかは，対象世界の背景知識に頼る部分が大きい．そのため，本論文では顧客の識別に重要な行動パターン (ルール) を自動的に抽出し，その「ルール」と「スコア値」を組み合わせて識別を行うダイナミックなシステムを，現象データマイニングシステムとして提案する．

現象データマイニングとは，ストリームデータと現象との関係を見つけるデータマイニングである．例えば，POS データはストリームデータ，顧客の振舞いは現象に相当し，現象データマイニングを行うことで，顧客の振舞いが分かる．これを現象データマイニングとしたのは，顧客の識別に重要な行動パターンをルールとして蓄積し，これを強力にすることで，顧客の振舞い (現象) を掴む事が出来るという考えのもとである．また，本論文では POS データの例を多く用いるが，提案する現象データマイニングは一般的にストリームデータと現象との関係を見つける手法として有効である．

2. 関 連 研 究

文献 [1] は，POS データから，顧客の年齢，性別，趣向など，顧客の特性を推測し，その特性をグルーピングすることで，顧客を特定する研究を行った．

本研究は顧客の振舞いを推測するデータマイニングであり，顧客を推測するものではない．さらに言えば，本研究では顧客の情報はすでに明らかである条件下で行うデータマイニングである．もし顧客が特定されていなければ，この文献の手法を用いることで顧客を特定し，本研究の現象データマイニングで顧客の振舞いを推測することが出来る．

また，文献 [2] は，相関ルールに繰り返し顧客購買率という顧客の振舞いを記述した指標を導入することより，頻出度がそれほど高くなくても顧客に繰り返し購買される重要なルールを発見する研究を行った．

本研究は，顧客の振舞いを正確に掴み，識別の精度を向上させ，さらにその識別されたクラス内に頻出する行動パターンを抽出する．頻出度がそれほど高くなくても，顧客に繰り返し現れるルールを発見することは，本システムにおいても有意であると考える．

3. 現象データマイニング

現象データマイニングは，蓄積されたストリームデータからは直接に分からない，それらのデータをもたらす現象との関係を見つけるデータマイニングである．

例えば，POS データはストリームデータ，顧客の振舞いは現象に相当し，現象データマイニングを行うことで顧客の振舞いが分かる．

現象データマイニングの目標は，識別に重要な顧客の行動パターンをルールとして自動的に抽出し，これをルールデータベースに蓄積する．さらに顧客の行動パターンの変化の度合いをスコア値としてルールのチェック項目に加え，ルールデータベースを強力にすることで，顧客の振舞い (現象) を掴むことである．その結果，ルールのみを用いた識別や，スコア値のみを用いた識別，さらには単に並行して用いるよりも高い識別精度を得ることができる．

3.1 システム構成

図 1 に，本論文で提案する現象データマイニングシステムを示す．現象データマイニングシステムは，過去のデータから識別器を作成し，識別された各クラス内で頻出するパターンをルールとして抽出する「学習部」と，実際のリアルデータを識別する「検証部」から成る．

学習部では，まず POS データから，次節で定義する顧客の振舞いデータを抽出する．振舞いデータは，顧客の行動パターンの変化を数値データで表現したものであり，次に，その顧客の振舞いデータを，カーネル法による SVM やロジスティック回帰を用いて識別し，顧客の行動パターンの変化の度合いを予測する識別器を作成する．さらに，分類・識別した各クラスで頻出する顧客の行動パターンをルールとして抽出し，ルールデータベースに蓄積する．

検証部では，まず学習部と同様に顧客の振舞いデータを抽出し，学習部で作成した識別器を用いて識別を行う．同時に，識別器により行動パターンの変化の度合いを出力し，データにスコアリングを行う．このスコア値と，振舞いデータをルールデータベース内のルールと照合することで識別を行う．

本システムは，リアルデータの識別精度によってルールの有効性を評価することができる．また，検証部においても，識別されたデータを蓄積することで，クラス内に頻出する最新の顧客の購買パターンをルールとして抽出し，ルールデータベースに加えることが出来る．つまり，学習部あるいは検証部で新たなルールが得られれば，それをリアルタイムで反映させることができるダイナミックなデータマイニングシステムである．さらに，このルールデータベースを強力にすることで，より顧客の振舞い (現象) を掴むことが出来る．

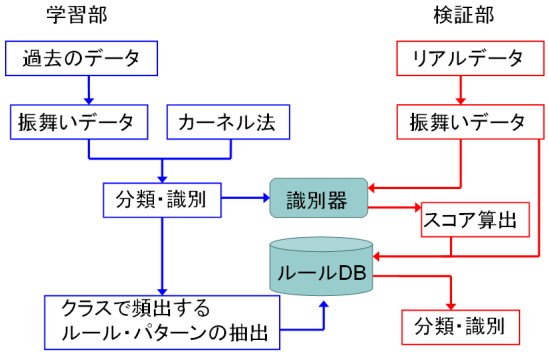

図 1　現象データマイニングシステム

3.2 振舞いデータ

本研究では，顧客の行動パターンの変化を数値データで表現する，振舞いデータを定義する．

POS データは，時間の経過と共に蓄積されるストリームデータである．このようなデータに対しマイニングを行うことで，

各顧客の行動パターン (購買パターン) が分かる．振舞いデータは,「顧客の行動が，過去のパターンと比較してどのような変化があるか」を示すものである．次に，振舞いデータの例をいくつか挙げる．

- 前回購入金額との差
- 前回購入金額との比率
- 前回購買日との間隔
- 前回購買商品との一致不一致
- 過去の平均の購買金額との差
- 過去の平均の購買金額との比率
- 過去の購入商品との一致不一致
- 過去の利用時間帯との一致不一致

図 2 は，ストリームデータに対して振舞いデータを付与し，これらを 1 つのリレーションとして用いることを示している．また，振舞いデータは，金額の変化，時間帯の変化，商品の変化など，あらゆる面から変化を監視することで，その結果，顧客の行動パターンの変化を考えることが出来る．POS データにあらゆる面から監視した振舞いデータを付与することによって，1 つのレコードは顧客の過去の行動パターンとの変化を示す．これを入力に用いることで，4.4 節で述べるカーネル法による分類・識別の精度を向上させ，顧客の行動パターンの変化の度合いを予測することが出来る．

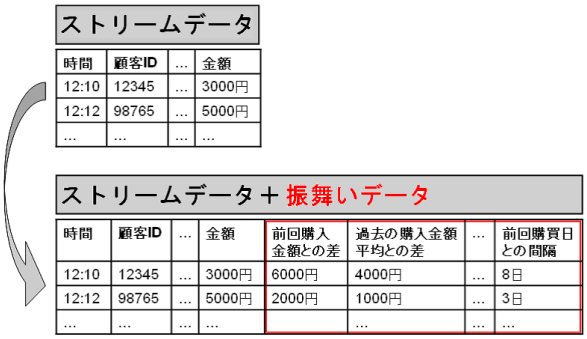

図 2　振舞いデータの付与

次に，振る舞いデータを定式化する考えとして、顧客分析の一つである RFM 分析を拡張したものを考える。RFM 分析とは、顧客について次の R、F、M、においてスコアリングすることで顧客のロイヤルティを推測する。

- R(recency： 最新購買日)
- F(frequency： 累計購買回数)
- M(monetary： 累計購買金額)

これらは、時間、頻度、金額という三つの属性を持ち、振舞いデータについても同様のアプローチで定式化する。

振舞いデータは，顧客の行動が時間上でどのように変化しているかを示したものである．つまり，どの期間で，どの属性が，どのように変化しているか，ということを考える必要がある．本論文ではこれらを時間軸，項目軸，比較演算軸として，三次元のリレーションとして振舞いデータを定式化する．まず，時間軸について考える．この時，どの期間の振舞いが識別精度の向上に有効であるかを考える必要がある．そこでまず本論文では，振舞いを考える期間として，特に L(last：前回) と P(period：期間) の二つが現象の推測に重要な情報であることに注目する．これらは,「前回の買い物と比較して，どのような変化があったか」という情報と「これまでの買い物と比較して，どのような変化があったか」という情報であり，RFM 分析における R(recency) を拡張したものである．次に，項目軸について考える．時間軸と同様，POS データに含まれるどの項目の振舞いが識別精度の向上に有効であるかを考える．この項目の選択には現象の背景知識が必要であるため，I(item：項目) という指標を用いる．この中には，RFM 分析における M(monetary) が含まれている．最後に，比較演算軸について考える．どのように振舞いを数値化すれば識別精度の向上に有効であるかを考え，Si(similarity：類似度) と St(statistic：統計処理) という指標を用いる．この Si は，一致不一致などを表現し，St は，差，比，頻度，傾向などを表現する指標である

- L(last：前回),P(period：期間)
- I(item：項目)
- Si(similarity：類似度),St(statistic：統計処理)

前述した振舞いデータの例は，これらの指標から導出したものである．

3.3　ルールデータベース

ルールとは，識別に重要な顧客の行動パターンである．これは，カーネル法によって識別された各クラス内で頻出する顧客の行動パターンを抽出することで得ることが出来る．例えば，メンバー会員登録している顧客に対して，カタログを渡すことで「購入する」クラスと,「購入しない」クラスに識別したとする．ここでもし,「購買する」クラス内に識別された顧客に,「先月，先々月も 3 万円以上購入している」というパターンが頻出していれば，これは「買う」か「買わないか」，つまりはカタログを渡すべきかを判断する重要なルールである．

このようなルールを抽出し，蓄積したものを本論文ではルールデータベースと定義する．ルールとして蓄積したルールデータベースは，いくつもの顧客の振舞い (現象) を記述したものであるといえる．つまり，これを強力にすることで，より正確に顧客の振舞いを掴むことが現象データマイニングの目標である．

図 3 は，リアルデータに対してルールを当てはめた例である．ルールデータベース内に「スコアが 500 以上，利用金額が 5 万円以上，前回との利用金額差が 10 万円以上ならばクラス A」というルールがある．これに対し，識別対象のデータは「スコア 500，利用金額 6 万円，前回購入金額との差 12 万円」であるため，スコア，利用金額，前回との利用金額差の全ての条件を満たしている．よって，このレコードはクラス A へ識別されることを示している．このように，ルールは以下の要素を条件としてチェックする．

- 現在の利用のチェック
- 過去の利用のチェック
- スコア値のチェック

現在の利用のチェックとは，リアルタイムで流れてくる現在の購買事実の項目に対するチェックである．例えば,「3 万円以

上の購買」や「商品カテゴリ A」などのようなルールを登録し，チェックする．

過去の利用チェックとは，その顧客の過去の購買事実に対するチェックである．例えば,「先月 3 万円以上の購入」や「1 週間以内で 3 回以上の来店」などのようなルールを登録し，チェックする．

スコア値のチェックとは，次節で詳細を述べるが顧客の行動パターンの変化をスコアリングし，そのスコア値に対するチェックであり，このチェック項目は，ルールを強力にする．顧客の行動パターンの変化が大きいということは，POS データの場合，その顧客の購買意欲が増加，あるいは減少していることに繋がり，何らかのプロモーションを行う必要があると分かる．またクレジットカードデータの場合，不正利用取引であることに繋がり，取引の中止を要求することが出来る．

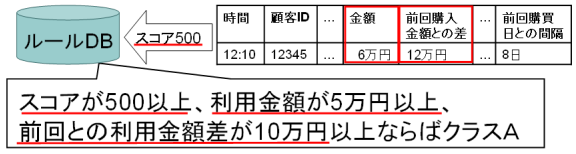

図 3　ルール当てはめ

3.4　スコアリング

顧客の行動パターンの変化は，顧客の購買意欲の増加・低下，異常行動，あるいは興味関心分野の移行を示している．前節で述べたように，この顧客の行動パターンの変化をスコアリングし，そのスコア値のみを用いて顧客を識別するのではなく，これをルール内のチェック項目に加え，ルールとスコア値を組み合わせて用いる．これにより，ルールのみを用いた識別や，スコア値のみを用いた識別，さらにはただ 2 つを並行して用いるよりも，高い識別精度を得ることができ，強力なルールを形成することが出来る．

正確なスコア値を予測するためには，精度の高い識別器を用いることが必要である．また，識別の精度が高ければ，識別されたクラス内で頻出するルールの信頼性もあるといえる．本研究ではカーネル法の理論を用いたロジスティック回帰や SVM を用いて識別器を作成する．振舞いデータを入力とすれば，これら 2 つのアルゴリズムは，顧客の行動パターンの変化の度合を連続量で予測する．この出力をスコア値として，ルールチェックの項目に導入する．

図 4 は，行動パターンの変化が少ない②に対し，①の方が高いスコア値を出力することを示している．この例は金額の変化の大きさを示したものであるが，実際には購入商品カテゴリの変化や，購入時間帯の変化など，あらゆる面の変化を監視し，振舞いデータを用いて顧客の行動パターンの変化を表現する．

本節で取り上げたロジスティック回帰や SVM は識別性能に優れたアルゴリズムであり，どちらが優れているとは一概に判断出来ない [11][12]．そこで，本実験ではクレジットカードのデータを対象に，ロジスティック回帰と SVM の性能を比較する．

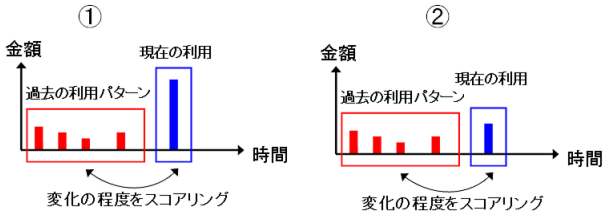

図 4　行動パターンの変化のスコアリング

3.5　カーネル法

POS データと振舞いデータを合わせたレコード一つ一つをベクトルとして表現することで，そのベクトルがどのクラスに属するかを考える．POS データは非線形構造のデータであるため，線形の識別・分類アルゴリズムを用いても，正確な結果を得ることは困難である．カーネル法を用いることで，この問題を解決することが出来る．

カーネル法は機械学習の一連の手法であり，本章ではカーネル法の概要と，数学的視点からとらえたカーネル法の性質であるカーネルトリックについて述べる．また，本実験で用いるカーネル法を用いた識別を行うアルゴリズムである SVM と，ロジスティック回帰についての簡単な説明を行う．

3.5.1　カーネル法の概要

高い識別性能を得るためには，対象の事前知識をうまく利用することが重要であり，カーネル法の特徴はこの事前知識を次のようなカーネル関数

$$k(x_i, x_j) \tag{1}$$

の形で表現することである．また，カーネル関数は二つの対象 (x_i, x_j) の類似度 R を定義したものであり，類似度という形で事前知識を表現し，学習に組み込むことが出来る．その関係式を式 2 に示す．

$$k(x_i, x_j) \to R \tag{2}$$

この類似度を定める数学的背景には，非線形構造データに対して高次元への写像を行い，線形識別を可能とするという理論がある．

通常，もとのデータ空間のデータ構造が線形関係であった場合，オブジェクト間の類似度は単に内積を考えることによって定まる．データ構造が非線形であった場合，図 5 で示すように高次元空間へ写像し，その写像関数の内積を考えることで類似度を定めることが出来る．つまり，この高次元空間で線形解析を用いることはもとの空間で非線形解析を行うことと等価となる．図 5 は (x_1, x_2) の空間を $(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ の空間に写像した図である．

カーネル法を用いることにより，ストリームデータが非線形なデータ構造であっても，線形解析手法を用いて分析を行うことが出来る．

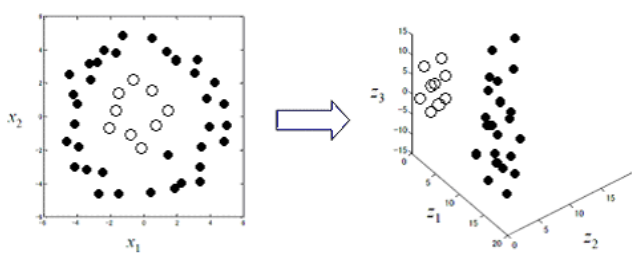

図 5　高次元写像の例

3.5.2　カーネルトリック

一般的に，分析対象となるサンプルの数が大きくなるほど，線形分離を行うことは難しくなる．また，高次元へ写像することでこの問題を解決しようとしても，サンプルの数と同程度の次元へ写像しなければならないため，サンプルの数が膨大であればその計算量も膨大なものになってしまう．カーネルトリックとは，この問題を解決する巧妙な仕掛けである．ある対象ベクトル x を，非線形の写像 $\Phi(x)$ によって元の空間 Ω から高次元空間 H へ変換し，その空間で線形識別を行うとする．ここで，もし高次元空間における二つの対象の内積が，式 3 のようにカーネル関数の形で計算できるなら，対象を高次元空間へ写像して内積をとるという計算を避けることが出来る．

$$\langle \Phi(x_i), \Phi(x_j) \rangle_H = k(x_i, x_j) \tag{3}$$

式 3 のように表せるカーネル関数を，正定値カーネルという．$k(x_i, x_j)$ が Ω 上の正定値カーネルであるとは，次の条件を満たすことである．

- 対称性

$$k(x_i, x_j) = k(x_j, x_i) \tag{4}$$

- 正定値性

任意の実数 $c_1, ..., c_n$ に対し，

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \tag{5}$$

また，上記の条件を満たす正定値カーネルには次のようなものが知られている．

- 多項式カーネル

$$k(x_i, x_j) = (x_i^T x_j + c)^d \tag{6}$$

- ガウスカーネル

$$k(x_i, x_j) = \exp\left(-\frac{1}{\sigma^2}\|x_i - x_j\|^2\right) \tag{7}$$

- シグモイドカーネル

$$k(x_i, x_j) = \tanh(a x_i^T x_j - b) \tag{8}$$

3.6　サポートベクターマシン (SVM)

SVM は，対象から抽出した特徴量を 1 つのベクトルとして考え，そのベクトル空間内でクラスを分類する超平面を引くことで，どちらに分類されるかを判断する．

パラメータの学習は，超平面と特徴ベクトルとの距離 (マージン) を最大にする「マージン最大化」という基準で行い，入力特徴ベクトル $\boldsymbol{x}$ に対して二値の出力値を計算する識別関数を作成する．

$$\boldsymbol{y = sign(w^T x - h)} \tag{9}$$

識別関数を導出する詳細な説明は避け，SVM の識別関数を以下に示す．

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{sign}(\mathbf{w}^{*T}\mathbf{x} - \boldsymbol{h}^*) \\ &= \boldsymbol{sign}(\sum_{i \in S} \boldsymbol{\alpha_i^* t_i t_j} \mathbf{x}_i^T \mathbf{x} - \boldsymbol{h}^*) \end{aligned} \tag{10}$$

また，非線形分離を可能にするカーネル学習を用いた SVM の識別関数を以下に示す．

$$\boldsymbol{y} = \boldsymbol{sign}(\sum_{i \in S} \boldsymbol{\alpha_i^* t_i t_j} \mathbf{x}_i^T \mathbf{x} - \boldsymbol{h}^*)$$

$$\iff \boldsymbol{y} = \boldsymbol{sign}(\sum_{i \in S} \boldsymbol{\alpha_i^* t_i t_j} \boldsymbol{\Phi}(\mathbf{x}_i)^T \boldsymbol{\Phi}(\mathbf{x}) - \boldsymbol{h}^*)) \tag{11}$$

$$= \boldsymbol{sign}(\sum_{i \in S} \boldsymbol{\alpha_i^* t_i t_j K}(\mathbf{x_i}, \mathbf{x}) - \boldsymbol{h}^*)) \tag{12}$$

3.6.1　ロジスティック回帰分析

ロジスティック回帰は，ある事象が発生する確率を予測する統計分析手法である．ある現象が発生する確率 $\boldsymbol{p}$ を，その現象の生起を説明するために抽出した $\boldsymbol{M}$ 個の特徴量からなる特徴ベクトル $\boldsymbol{x^T = (x_1, ..., x_M)}$ で説明しようと考える場合，$\boldsymbol{x^T = (x_1, ..., x_M)}$ という状態のもとで現象が発生するという条件付確率を $\boldsymbol{p(x)}$ で表す．ここで，$\boldsymbol{M}$ 個の特徴量の影響を，線形な合成関数

$$\boldsymbol{Z = \beta_0 + \beta_1 x_1 + \cdots + \beta_M x_M} \tag{13}$$

で表現し，この合成関数 $\boldsymbol{Z}$ にロジスティック関数

$$\boldsymbol{p(x)} = \frac{\boldsymbol{\exp(z)}}{\boldsymbol{1 + \exp(Z)}} = \frac{\boldsymbol{1}}{\boldsymbol{1 + \exp(-Z)}} \tag{14}$$

を用いたものが，ロジスティク回帰モデルである．

$$\boldsymbol{p(x)} = \frac{\boldsymbol{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_M x_M)}}{\boldsymbol{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_M x_M)}} \tag{15}$$

ロジスティック回帰の学習法は最尤法であり，各パラメータの微分値を用いて，少しずつパラメータを更新していく．そこで，この更新の度に加えられる各パラメータの微分値を

$$\boldsymbol{\Phi(x^{(j)}) = \sum_{i=1}^{N} (u_i - y_i) x_{ij}} \tag{16}$$

とすると，パラメータベクトル $\boldsymbol{\beta}$ は，更新のたびに $\boldsymbol{\Phi(x^{(j)})}$ が加えられているか引かれているかのどちらかである．そこで，

線形和によってパラメータベクトルは

$$\beta = \sum_{j=1}^{N} \alpha \Phi(x^{(j)}) \tag{17}$$

と表現できる．これをロジスティック回帰モデルに代入すると，

$$\log\left[\frac{p(\mathrm{x})}{1-p(\mathrm{x})}\right] = \langle \beta, \Phi(x) \rangle = \sum_{j=1}^{N} \alpha \langle \Phi(x^{(j)}), \Phi(x) \rangle \tag{18}$$

というカーネル関数を用いた回帰式が得られる．

4. 実験と評価

4.1 目　　的

これまでに述べた現象データマイニングは，POS データと顧客の振舞いの例を多く出した．POS データと同様にストリームデータであるクレジットカードの取引データに対しても，現象データマイニングは有効である．この場合，クレジットカードの取引データはストリームデータ，カードを利用する顧客の振舞いは現象に相当する．

最終的な目標は，クレジットカードの取引データを不正利用クラスと正常利用クラスとに識別し，不正利用クラス内で頻出する顧客の行動パターンをルールとして蓄積し，これを強力にすることであるが，本実験ではそこまでに至らず，不正利用と正常利用に識別するまでの実験を行う．

識別アルゴリズムは，SVM とロジスティック回帰を用いた．これら 2 つは識別性能に優れたアルゴリズムであり，どちらが優れているとは一概に判断出来ない．そのため，本実験でその精度を比較・考察する．

4.2 実験の方法

以下に，使用した実験データの内容を示す．

- データの属性数
  - 取引データの 16 属性と振舞いデータの 39 属性の計 55 属性
- サンプリング件数
  - モデルの作成用：30 万件　(内不正利用件数 300 件)
  - 検証用：20 万件　(内不正利用件数 1000 件)

通常，実際のクレジットカードの取引データの属性数は 100 近く存在する．しかし，不正利用に結びつかないような情報 (顧客 ID や生年月日など) は分析対象にせず，16 属性を選択した．これにさらに振舞いデータの 39 属性を加え，計 55 属性のデータ項目を用いる．以下に，その 39 属性の振舞いデータの内，例として 3 属性を挙げる．

- 前回利用金額との差
- 前回加盟店コードとの一致不一致
- 過去 6 ヶ月の曜日ごとの利用回数との比較

よって，一件の取引データに対して 55 属性のデータ項目が存在することとなる．前述した SVM とロジスティック回帰は，これら 55 属性のデータ項目から成る取引データを一つのベクトルとして捉え，その取引データが正常利用か，不正利用かを識別する．

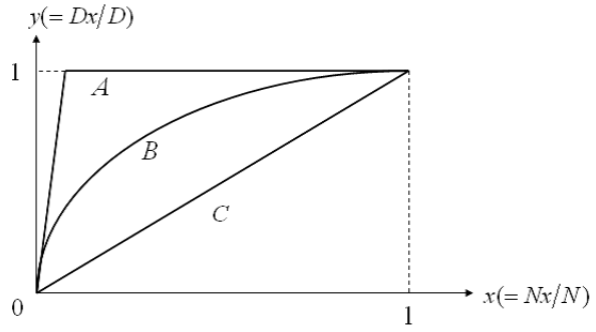

図 6　CAP 図の評価

識別精度が高ければ，識別されたクラス内で頻出するルールの信頼性は高いといえる．また，スコアリング精度が良ければ，ルールと照合した時に適正なクラスへ識別することが出来る．スコアリングの評価には CAP 図，識別精度の評価には検知率・誤認率を用いる．図 7 は，CAP 図，検知率，誤認率を求めるために必要なデータと導出方法を，図で表したものである．

CAP 図とは，モデルの識別性能を表す図である．取引データの総件数を N として，そのうち実際に不正利用であった件数を D とする．各取引データで，SVM またはロジスティック回帰によって予測したスコア値を高いほうから順に並べ，スコア値が高いほうから N_x とってきた場合に，その中に含まれる実際の不正利用の件数を Dx とすると，$x = \frac{N_x}{N}, y = \frac{D_x}{D}$ として描いたグラフが CAP 図である (図 6).

モデルに全く説明力がない場合，予測した確率に関係なくランダムに不正利用のデータが含まれているため，どこをとっても不正利用のデータが含まれる確率は同じとなり，グラフは図 6 の C の線 $x = y$(45 度線) となる．

また，理想的なモデルであれば，実際に不正利用であったデータは高い確率で予測できるため，グラフは図 6 の A の線となる．

実際には両者の中間をとる B のような曲線となり，この曲線が A に近いほど判別能力の高いモデルであり，C に近づけば判別能力のないモデルであるといえる．よって，この曲線が A に近いほど，実際の不正に対して現象データマイニング

検知率とは，取引データに存在する全ての不正利用の取引の件数 S の内，不正利用であると判別できた件数 S_x の割合であり，$\frac{S_x}{S}$ で示す．

誤認率とは，不正利用であると判別された取引の件数 T の内，実際には不正利用ではなかった件数 T_x の割合であり，$\frac{T_x}{T}$ で示す．

4.3 結　　果

4.3.1 SVM モデル

前節の図 6 と図 7 で CAP 図評価方法，および検出率・誤認率の導出方法を述べたが，図 8 は SVM によるスコアリングを行い，そのスコア値を高い方から順に並べたものが横軸であり，

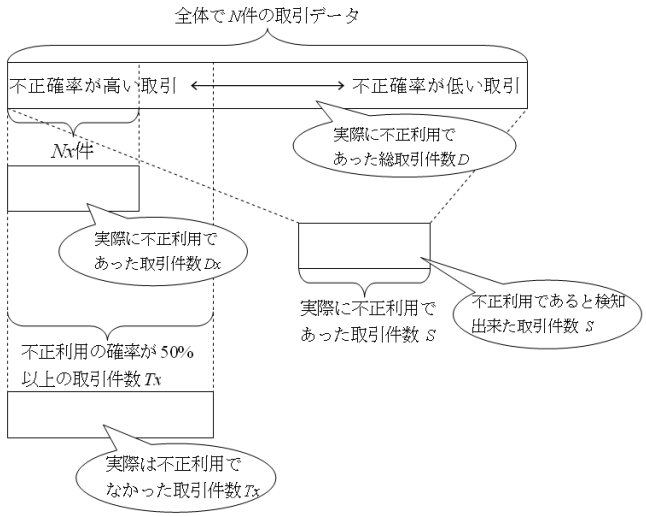

図 7　評価算出の方法

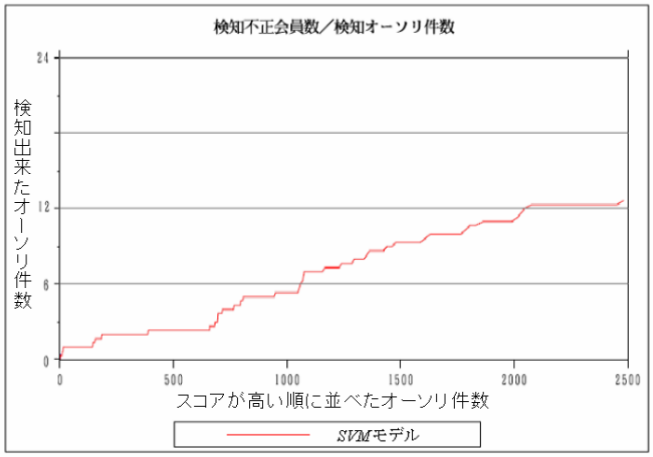

図 8　SVM による CAP 図 1

ヒットオーソリ数		ヒット会員数		検知率		誤認率	
ヒット全件	ヒット不正	ヒット全件	ヒット不正	オーソリ	会員	オーソリ	会員
3,655	56	2,995	22	5.86%	14.28%	98.46%	99.26%

図 9　SVM による識別性能

その中で実際に不正であった顧客の人数を縦軸で表している．よって，図 6 の ***A*** に近い方が，高いスコアで多くの不正を検知出来ており，スコアリング精度の高いモデルであるといえる

また，図 9 は取引データ (オーソリ) 基準の識別率・誤認率と，顧客基準の識別率・誤認率を示している．

4.3.2　ロジスティック回帰モデル

前節の SVM の時と同様，図 10 はロジスティック回帰によるスコアリングを行い，そのスコア値を高い方から順に並べたものが横軸であり，その中で実際に不正であった顧客の人数を縦軸で表している．図 6 の ***A*** に近い方が，高いスコアで多くの不正を検知出来ており，スコアリング精度の高いモデルであるといえる．SVM と比べて，ロジスティック回帰の方が図 6 の ***A*** に近い曲線を描いており，ロジスティック回帰の方がスコアリング精度の高いモデルであるといえる．

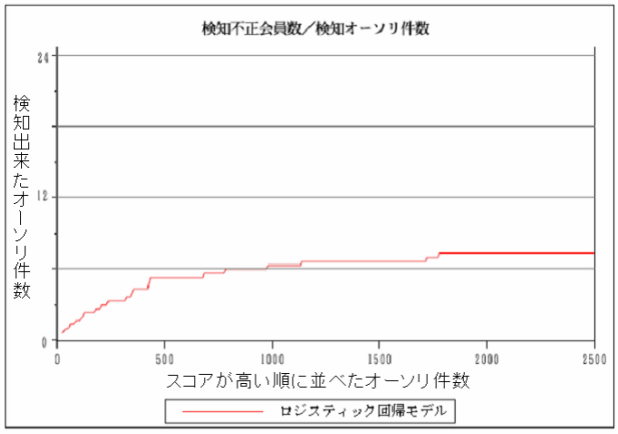

図 10　ロジスティック回帰による CAP 図

ヒットオーソリ数		ヒット会員数		検知率		誤認率	
ヒット全件	ヒット不正	ヒット全件	ヒット不正	オーソリ	会員	オーソリ	会員
13,838	220	7,889	29	23.03%	18.83%	98.41%	99.63%

図 11　ロジスティック回帰による識別性能

また，前節と同様に図 11 は取引データ (オーソリ) 基準の識別率・誤認率と，顧客基準の識別率・誤認率を示している．

5.　考　　察

CAP 図はスコアリングの精度を示し，検知率・誤認率は識別の精度を示す．結果として，ロジスティック回帰は不正利用の取引データを高いスコア値で検知することには長けており，学習・当てはめの処理でも高速に処理することが出来た．しかし，不正利用取引の検知率・誤認率を比較すると，SVM と性能はほぼ変わらないという出力結果を得た．SVM は最適なパラメータを発見するために実験を試行錯誤で行わなければならず，今後調整していくことで判別の精度が向上することが出来ると考える，しかし一方で，学習・当てはめに時間がかかり，また，パラメータの少しの変化でモデルの精度が大きく変化してしまうなどの問題もあり，今回の実験のように大規模なデータが対象であった場合，処理効率・スコアリング精度の面ではロジスティックが優り，識別精度の面では僅かに SVM が優るという結論を得た．結論として，現象データマイニングのスコアリングにはロジスティック回帰を用いた方が良いと判断できる．

5.1　ロジスティック回帰での検知

CAP 図から考察できることは，ロジスティック回帰は不正利用の取引データを高いスコア値で検知することが出来たという点である．さらに，ロジスティック回帰は，実行処理が短時間ですむ．これは，実際にビジネスに活用する場合に非常に重要な点である．本研究の実験では一番バランスの良いモデルであるといえる．

5.2　SVM での検知

実験はシグモイドカーネルとガウスカーネルを用いた SVM で行った．SVM の場合，予測するスコア値はマージンそのものである．そのため，CAP 図を見て分かるように，不正に対し

て高いスコアを出力できていなく，予測スコア値が低くなっても検知の精度を保っている．これは，明らかな不正利用であったとしても，SVM は予測スコア値を高スコアで出力することが出来ないことを示している．SVM は異なるクラス間のマージンを最大にするよう分離超平面を引くため，学習用データ内に不正利用のサンプルが少ない場合，不正利用のサンプルのほとんどがサポートベクターとなってしまう．これが，SVM が高いスコアを出力できない理由である．

また，SVM はカーネル関数のパラメータを少し変更するだけで，モデルの特徴は大きく変わってしまう．さらに学習・あてはめサンプルの量が非常に多い場合，その処理に多くの時間がかかってしまう．また，SVM において最適なパラメータは試行錯誤で求めていくしかないため，大規模なデータが対象に識別能力の高いモデルを作成しようとすればさらに大量の時間ががかると予想する．

6. おわりに

本論文では，顧客の振舞いが分かる現象データマイニングを提案した．そのシステムは，POS から振舞いデータを作成し，カーネル法によって顧客の振舞いの変化を識別・スコアリングする．その識別された各クラス内で頻出する顧客の行動パターンを，ルールとして自動的に抽出し，ルールデータベースに蓄積する．さらに，リアルタイムで新たなルールの導入，ルールの有効性の評価，顧客の購買行動の変化をスコアリングしてルールのチェック項目に導入することで，そのルールデータベースを強力なものにし，顧客の振舞いを掴む．

本実験ではカーネル法によって顧客の振舞いの変化を識別・スコアリングを行った．これは識別されたクラスから抽出するルール信頼性，さらにはルールデータベースそのものの信頼性に繋がる．また，振舞いデータを 36 項目用いたが，顧客の行動パターンの変化をより良く説明する振舞いデータの抽出は，識別・スコアリング精度の向上，さらにはルールデータベースの強化に繋がる本システムの重要な部分である．識別アルゴリズムにはロジスティック回帰と SVM を用いて，CAP 図，検出率，誤認率，処理時間に対して比較を行った．その結果，スコアリング精度，処理時間の面でロジスティック回帰が大きく上回り，本システムにおいてはロジスティック回帰の方が優位であると判断した．

今後の課題としては，自動的に抽出したルールをルールデータベース内に蓄積し，リアルタイムで新たなルールの導入，ルールの有効性の評価を行い，そのルールデータベースを強力なものにするダイナミックなシステムの構築が挙げられる．

謝　辞

本研究の実験で使用したクレジットカードのデータの提供，及び実験に対するアドバイスを頂いた (株) インテリジェントウェイブの関係者方々に御礼申し上げます．

文　献

[1] John McCarthy, PHENOMEMAL DATA MINING : FROM DATA TO PHENOMENA, SIGKKD Explorations, vol.1, no.2, pp.24-29, Jan 2000.

[2] 裴明花，谷口伸一，原隆浩，西尾章治郎”重要な顧客層および相関ルール発見のための繰り返し購買パターンを考慮した相関ルールマイニング，”情報処理学会論文誌，vol.47，No12，3352，Dec.2006.

[3] Jing Wu, Zheng Lin, ”Research on Customer Segmentation Model by Clustering,”Proc 7th international conference on Electronic commerce, no.39, pp.316-318, Xi’an, China, August 2005.

[4] Tadahiko Sato, Tomoyuki Higuchi, Genshiro Kitagawa, Statistical Inference Using Stochastic Switching Models for the Discrimination of Unobserved Display Promotion from POS Data, Proc. Marketing Letters, vol.15, no.1, pp.37-60, October.2004.

[5] 麻生英樹 津田宏治 村田昇，統計科学のフロンティア 6 パターン認識と学習の統計学, 大塚信一（編）,（株）岩波書店，東京，2003.

[6] 赤穂昭太郎，”カーネルマシン，
” http://www.neurosci.aist.go.jp/ akaho/papers/kernel-akaho.pdf.

[7] 鹿島久嗣，”カーネル法による構造データマイニング，”, 情報処理，vol.46，no1，pp27-33， Jan.2005.

[8] 栗田多喜夫，”顔検出・顔認識のための統計的手法，” http://www.neurosci.aist.go.jp/ kurita/lecture/statface.pdf, November.2002

[9] 丹後俊郎 山岡和枝 高木春良，ロジスティック回帰分析-sas を利用した統計解析の実際-，朝倉邦造（編）,（株）朝倉書店，東京，1996.

[10] 芳賀敏郎 野澤昌弘 岸本淳司，SAS による回帰分析，五味文彦（編），東京大学出版会，東京，2002.

[11] S. C. Hoi, R. Jin, and M. R. Lyu, ”Large-scale text categorization by batch mode active learning,” Proc. 15th International Conference on World Wide Web, pp633-642, Edinburgh, Scotland, May.2006.

[12] J. Zhang, R. Jin, Y. Yang, and A. Hauptmann, ”Modified logistic regression: An approximation to svm and its applications in large-scale text categorization. ”Proc 20th International Conference on Machine Learning (ICML), Washington, DC, USA, August 2003.

[13] Paul Komarek, Andrew W. Moore, ”Making Logistic Regression a Core Data Mining Tool with TR-IRLS,”Proc 5th IEEE International Conference on Data Mining, pp.685-688, DC, USA, November 2005.

[14] Steven H. Low, Nicholas F. Maxemchuk, Sanjoy Paul, ”Anonymous Credit Cards and Their Collusion Analysis,” Proc. IEEE/ACM, vol.4, no.6, pp.809-815, December 1996.