インターネット・マーケティングと遺伝的プログラミングによる実現ツールの開発

高木, 昇
九州産業大学・商学部・商学科

時永, 祥三
九州大学大学院経済学研究院

https://doi.org/10.15017/7640

出版情報：經濟學研究. 72 (4), pp.61-89, 2006-03-31. 九州大学経済学会
バージョン：
権利関係：

インターネット・マーケティングと遺伝的プログラミングによる実現ツールの開発

高木 昇　　時永 祥三

1　まえがき

商品を製造し, あるいはサービスを商品化し, 市場で販売するという, 企業の基本的な行為に, マーケティングという分野が加わったのは, 歴史的に新しいものではない。多量生産システムにより, 多くの大衆が消費者として登場して以降, 過剰生産と販路の確保, 他社との競合は, 避けて通れないものとなった。現在, そのマーケティングの媒体として, インターネットの果たす役割が重視されるようになっており, インターネット・マーケティングの概念も整備されつつある。

本論文では, このようなインターネット・マーケティングの現状と, 課題について述べるとともに, われわれが実施している, インターネット・マーケティングと遺伝的プログラミングによる実現ツールの開発について示す。

Kotler により分析された 3 つの基本的な段階とは, 生産, 販売, ブランド管理であり, これにより新しく出現した消費市場を分析することが行われてきた。しかしながら, この段階における製品管理においても, 消費者からの視点よりは, 生産のブランド化製品の生産個別化が強調されていた。現在にいたるこの第 4 段階は, 顧客管理の段階として位置づけられている。本論文では, インターネットを用いたマーケティング, パーソナライゼーション, ブランド形成のステージ管理, 新製品開発とモジュール化の方法論, エクストラネット構築, コミュニティの形成, 価格設定, ウェブによる販売チャネルについて, 現状と課題を整理し, インターネット・マーケティングの将来を展望する。

論文の後半では, われわれの提案する遺伝的プログラミング (Genetic Programming:GP) の手法に基づく顧客管理・文書管理のシステムを述べる。具体的には, 顧客の嗜好を分類し管理する場合に, 顧客の属性を入力として購買するであろう商品を推定する方法を GP によるルール生成により実施する。この基本的な方法論は,GP による学習を用いており, コンパクトで, 効率的な顧客に関する情報の整理が可能となる。顧客のクラスタ分類・検出に関して, これまでの研究では多変量解析によるクラスタ分析, 多次元尺度法やニューラルネットワーク法が提案されているが, 用いる変数は数値型変数に限定され, 数値的な判別精度が改善されても, これがユーザに分かる言語の形で提供されない問題がある。言語的なルールによる判断が出力されることにより, 分類の根拠が解釈可能な形で示されるメリットがある [4][5]。また,ID3 などの従来の演繹推論の手法では, 数値データとカテゴリデータとが, 同時に 1 つの大きな判別の木構造の形で示されるため, ルールの構造が複雑となる [6]。本論文では GP によるルール推定の手法を用いて, 顧客情報を管理するシステムを提案し, 実際のデータに応用する。

また, クラスタ分類の問題と同時にある顧客クラスタの特徴を記述する問題がある。これをクラスタ特徴記述の問題と呼ぶことにする。この手法を開発する場合に, 演繹推論を分類手法として当初の目的から拡張して用いることは, ペアサンプルと呼ばれる対立的 (一方が合格なら他のグループは不合格) であるデータ集合が準備できないなどの理由から困難である。従って, 検索されるべき対象があらかじめ学習データとして与えられてない場合に相当するクラスタ検索については, 別の方法を適用する必要がある。本論文では,GP 手法をクラスタ分類手法をクラスタ特徴記述へと拡張している。具体的には, 同時に, 特定の顧客の商品嗜好を推論し抽出するため, データの集合から特徴的なクラスタを検出する手法を提案し, その応用について述べる。GP 手法の基本は, カテゴリ化された変数により記述されるデータ集合に対して, 論理演算を実施する木構造 (GP における個体に相当する) を多数与えおき, 適合度に応じて遺伝的操作を, 安定的に検出されるクラスタが発見されるまで繰り返し, これにより最終的にクラスタを特徴付ける論理式を求める方法である。

最後の章では,GP 手法に基づいた文書分類・検索システムについて述べるインターネット・マーケティングの環境においては,顧客からの意見の集約や,さまざまに収集された文書を,管理することが必要になる。文書分類の手法として,これまで形態素解析などの自然言語処理により単語をキーワードとして抽出し,これらの出現頻度をもとにした特徴ベクトルを求め,この特徴ベクトルに関するクラスタ (特定の特徴をもった文書の集合) 重心の計算と,分類するべき文書の属性との距離の計算を用いる方法が用いられている。しかしながら特徴ベクトルによる分類方法では,キーワードの出現頻度だけが重視され,キーワード間の出現順序が無視されている問題があり,単純なベクトル空間の距離ではなく,一般化された関数として文書間の距離を定義することが望ましい。本論文では,GP 手法を用いて文書分類を実現するシステムを提案し,実際のデータに応用する。

2　マーケティング概念の変遷

2.1　伝統的なマーケティング手法

資本主義の典型である米国においてさえ,20 世紀初頭には,ほとんどの人口は農村にくらしており,現状とはかけ離れた状況にあった。しかし,産業革命による新しい動力の出現と,多量の輸送機関の整備は,生産地における多量生産と,これを大規模な消費地に送りとどける,生産と流通のパターンを形成することになる。

20 世紀初頭に経済活動を支配した概念は,規模の経済であり,多量生産である。1913 年に開始されたフォード式乗用車の生産方式は,その原点とも言えるであろう。多量生産は,同時に,標準化の必要性をともない,これによりコスト削減が実現されることとなる。

このような規模の経済の拡大にともない,製品をできるだけ広い範囲に流通させることが,マーケティングの役割である。マーケティング理論の創始者である Kotler は,これを生産概念とよび,「企業志向型組織の管理者は,生産の高い効率性を達成し,同時にできるだけ広い範囲に流通させることを目的とする」として位置づけている。

しかしながら,大規模生産による多量の製品の製造と,流通による広域の交易は,すぐに限界に達することが認識される。そのステージは,いつであるかは明確にはされないが,1970 年代には,すでにその段階に入っているとする意見がある。やがて,製品の効率的な生産と流通だけでは説明できない社会が出現することになり,これを発展段階として解明することがなされてきた。

Kotler により分析された 3 つの基本的な段階とは,生産,販売,ブランド管理であり,これにより新しく出現した消費市場を分析することが行われてきた。

第 1 段階の「生産」の効率化においては,極めて整備された生産方式のもとで,製品の大幅なコストダウンが実現され,多くの生活必需品が,広範囲に低価格で提供されることとなる。しかし,同時に購買力を高めた消費者は,実質本位の使用価値だけの製品から,更に進み,魅力的な商品を求めることになる。これにより出現するのが,第 2 段階である「販売」である。この段階における販売概念での基本では,消費者に何も働きかけない企業の製品は十分に売れることはないだろうし,そのための攻撃的な販売とプロモーションが必要である。商品は,やがて標準的なものから,固有の製品ラインをもった個別的な,特徴あるものとして製造される。また,市場規模も,全国にまで拡大し,この広範囲な市場で通用するブランドをもつことが追求されてきた。

第 3 段階であるブランド管理の概念のもとでは,消費者が求める満足感を,競合他社より効率的に,すばやく実現することが基本とされてきた。ラジオやテレビなどの新しいコマーシャル媒体の出現は,これを加速することになり,他社に先駆けて製品のブランド確立をはかることに,多くの努力が払われることとなる。

しかしながら,この段階における製品管理においても,消費者からの視点よりは,生産のブランド化製品の生産個別化が強調されていたことも分析されている。すなわち,消費者の満足感は,製品本位の優越さにより代表されると考えられてきた。

この限界を取り払うものとして投入されたものがコンピュータであり,特に消費者である顧客の情報を,有効に管理し,活用することの重要性に目が向けられることとなる。その延長上に,現在のマーケティングが議論されていると言える。

現在にいたるこの第 4 段階は,平凡な言葉ではあるが,顧客管理の段階として位置づけられている。1960 年代に

商用が開始されたコンピュータの利用分野として, 早くから顧客データの管理, すなわちデータベース化が志向されている。しかしながら, コンピュータの容量は極めて限定され, しかもコストが高いなどの問題点は, この基本的な構想を実現するには大きな障害であった。また, データの活用に関しても, とりあえず蓄積することに主眼がおかれ, これを分析するツールを整備するまでの余力はなかったと言える。

しかし,1980 年代に入り, 顧客データベースを利用した郵便物の自動送付のシステム, いわゆるダイレクトマーケティングの方法論が導入されることとなる。これにともなって, カタログ販売や個人へのプロモーション, あるいは, クレジットカードなどの決済手段との結合など, より現実的な展開を見せ始める。

以上, 概括したように, 現在のマーケティングの到達点は, 顧客データベースの管理と, 顧客への個別対応として整理できるであろう。個別的な製品やサービスの提供, 単なるブランド管理ではなく, 企業が示すコミュニティへの参加意識の醸成などを支援するシステム構築が求められている。

2.2 インターネットマーケティングの特質

現在では, 企業の内部システム管理に関連して, いわゆる専用線による接続形態が残されているが, 情報ネットワークの大半は, インターネットによる通信へと大幅に変換されている。専用線はデータのセキュリティ管理の面からは望ましいが, コストがインターネット利用に比べて高額であることや, 相手先ごとに設定する必要などがあり, マーケティングなどの, 広範なユーザを対象とする業務には向かない。

更に, インターネットがマーケティングに果たす役割は, これにとどまらず, 広く意見を求める場合の簡単さや, ユーザにおける自主的なグループを形成できる可能性などの機能が加わる。このような背景から, 現在では, マーケティングにおけるインターネットの価値は大きくなり, インターネット・マーケティングとよばれる分野が出現している。

以下では, インターネット・マーケティングを実施する場合に, ポイントとされていることをまとめる。なお, 本論文ではこれらのすべてを記述することは適切ではないので, いくつかの限られたポイントだけを, 後で詳述する。

インターネットを用いたマーケティング

すでに述べたように, 専用線からインターネットへの移行は, 企業内システムを含めて順次実施されており, 特に, 一般の消費者を対象として構成される情報ネットワークは, 現在ではインターネットである。インターネットの活用範囲の拡大とならんで, そのセキュリティ側面の強化などの必要性が説かれてはいるが, 専用線の時代に逆戻りすることはない。

インターネットの特質として利用コストが安価であることや利用のためのプロトコルが共通化されその面から簡便であることが強調できる。更に, この特質からくるオープン性の大きな要因となっている。例えば, 電子調達の分野においても, 企業の独自の調達サイトを維持管理するのは極めて困難であることが証明される一方で, いわゆるマーケットプレイスとしてオープン化されたサイトは, その機能をより良く発揮することができる。

企業は商品のプロモーションや, これに関連した企業の宣伝の媒体として, いちはやくインターネットに注目し, サイトを立ち上げたが, 初期の段階では企業イメージを植えつけることに主眼が置かれ, 商品のセールスや, マーケティングに活用することは検証されてこなかったと言える。

別の側面として, いわゆるドットコム企業の株式をめぐる大きな混乱が存在した。投機的な投資家は, ドットコム企業の企業業績に注目するのではなく, その株式の値上がりにだけ期待して投資を行う傾向が顕在化し, アマゾンなどの代表的な企業に対するイメージも損なわれることになった。また, 従来の流通チャネルとの競合や, インターネット・マーケティングの特質は何かが問われることとなった。

しかし, インターネットによる販売が本格化する従って, その効果は伝播的に拡大していった。例えば, ルーター製造の大手メーカである cisco 社では, 装置の不具合からユーザからの質問やこれに対する回答をインターネットを介して実施したが, この副次的な効果として, ユーザがお互いに意見を交換する場が形成され, これが更に商品販売に有利に作用することになった。これは, 企業間でも効果として現れてきており, インターネットによる商品販売は, いわゆる検索サイトやサービスを拡大することになり, オークションなどの形式を構成させる効果を生んで

いる。

更に,音声や画像を効率的にインターネットで配信できる技術が確立されたことも,インターネット・マーケティングを促進する要因となっている。いわゆる,ブロードバンドによる多量の同時配信は,多くのユーザに,安価で最新のデータを視覚的に得ることができる。

パーソナライゼーション

ブランド管理から顧客管理への流れは,インターネット・マーケティングに先立つ段階で認識されてはいたが,現在では,より強く意識されている。ブランド管理は,いわば,優良な製品であることを企業が消費者に強調することに力点が置かれるが,パーソナライゼーションの段階においては,消費者の理解へ力点が移行する。多くの事例がとりあげられているが,例えば新聞社が特定の顧客に対して,特別に収集した最新のニュースを配信するなどのサービスはよく知られている。

このような,商品やサービスの高度なカスタマイズ化個別化により,競合他社との差異を際立たせることが必要とされている。これを実現する方法論として,すでに述べた,顧客データベースやデータマイニング手法が適用される。

ブランド形成のステージ管理

ブランド管理は,いわば1つ前の段階であると言えるが,しかしながら,インターネット・マーケティングの段階においても,大きな役割を果たすことになる。具体的には,インターネットを通じて多量の情報が消費者に流れる状況では,消費者の選択が,逆に狭まる可能性が存在することを示唆している。あるいは,インターネットによる広範囲なマーケティングの機会の拡大は,新規の企業の参入の障壁を低くしているように思えるが,実際には参入は容易でないケースが多いことも反映している。例として,よくあげられる英国でのインターネット銀行の立ち上げと失敗がある。スタート時点では,多くの注目を集め順調に推移したが,コスト高から予定したサービスを実施できずに,行き詰る結果となった。この企業の金融事業へのノウハウの欠如が大きな原因であるが,背景には,信頼できる企業かどうかを見極める消費者の存在がある。ブランド形成に有利である企業は,実は,インターネット以前の段階でも,ユーザに受け入れられている企業であるケースが少なくない。更に,情報があふれる時代においては,ユーザは選択する基準として,企業の安定性,すなわちブランド力に依存することが多くなる。

新製品開発とモジュール化の方法論

製品製造が計画されてから市場へ投入され,やがては市場から消えていく,いわゆる商品のライフサイクルが,従来に比べて短くなっていることが指摘されている。市場は,常に新しい商品を求めているが,その存在価値は短くなっている。これにともなって,より効率的に商品を開発し,製造することが必要となっている。

製品開発で重要なポイントは,消費者のニーズの把握と,設計ミスの排除であると言われている。これらは相互に関連しており,インターネット・マーケティングにおては,広範な消費者のニーズを,調査や統計解析により求めるとともに,新しい素材に関する情報などを集積して,果たした商品として採算がとれるかが検証される。また,いったん市場にでた場合の重要な要素として,標準化がある。ある企業のデータでは,数百にのぼる当社の製品の中で,収益に貢献しているのは数個の製品であり,そのすべてが,市場で「標準品」として通用しているものであるとされている。

また,ある企業の例では,2年以内の販売された製品の利益への貢献度は約8割であり,多くの製品のラインナップの中で重要なものは,数個に過ぎないことが理解されている。そのため,製品の開発の速度が重要となる。このような迅速な製品の開発と製造を実現する方法論として,現在,注目されているものがモジュール生産,あるいは生産システムのモジュール化である。この基本は,部分的に優れた技術や製品をもった企業との連携により,より早期に製品を製造する,そのためには自社の部品や技術にはこだわらないことが基本となる。また,製造過程を含めて,分割可能なシステムにすること,これによる管理を簡素化し,効率化することが必要となる。電子調達などのシステムを用いて,より良好なパートナーを探し,関係を強化することが行われる。

エクストラネット構築

ネットワークの構成分類には,パブリック・インターネット,イントラネットおよびエクストラネットの3つが存在する。これらは,どれも同じインターネットを基盤として構築されるが,ユーザがアクセスできる範囲や,セキュリティ管理の種類を変更することにより,目的別の構成となっている。パブリック・インターネットは,すべての

ユーザに公開される情報のサイトにより構成されるものであり, 一般的なマーケティングはこれにより行われる。

イントラネットは, 企業の社員あるいは, その中でも限定された社員にアクセスが許されたサイトから構成されるシステムであり, 社員管理役員からのメッセージのほかに, 販売支援システムなども, イントラネットを通じてアクセスが行われる。イントラネットの実現や管理には, ファイヤーウォールによる技術で行われる。

このような, 広範に開かれたネット, あるいは限定された範囲のネットの中間に位置するものが, エクストラネットである。これは, 社外からの社内情報への限定的なアクセスを許す方法であり, 現在までの実績では, さまざまな企業間の取引の電子化や, 契約の実行に有効であることが示されている。例えば, ある製造業の企業が, 顧客である企業から修理のための部品の注文を受ける場合を想定する。従来の方式では, この注文を受け付ける要員を配置し, これを再度入力することにより, 部品の注文が完結し, 更に配送なども確認する必要があった。しかし, エクストラネットのもとでは, 顧客は当該の企業の在庫ファイルにアクセスし, その部品の存在を自身で確認して, 更にその時点で注文をだすことができる。部品の配送に関しても, 荷物の追跡システムがエクストラネットとして構成され, 顧客自身でその配送の現状を確認することができる。

以上のように, 企業間の電子商取引 (Business to Business:B2B transaction), あるいは電子調達の業務にエクストラネットは大きな役割を果たしている。

コミュニティの形成

インターネットを通じて意見を交換する仕組みは, どこでも実現可能であるが, インターネット・コミュニティ, あるいはオンライン・コミュニティとよばれるものは, これらの仕組みの中でも, 企業活動と密接に結びついたものをさしている。インターネットにおける, いわゆるチャットによる意見交換はその一部であるが, オンライン・コミュニティで想定されているものは, 企業の販売する製品や, 企業そのものへの意見を交換する場である。

インターネット・マーケティングの立場から言えば, このようなオンライン・コミュニティは, 消費者の意見を収集するのに最適な場所であると言える。しかし, 一方では, 消費者からの厳しい意見が掲載され, これを広範囲に公開されるというリスクも含まれている。例えば, 化粧品について消費者から製品に関するさまざまな意見が寄せられ, これを参考にして, より望ましい商品の選択ができるであろう。しかし, 同時に, 悪意をもった消費者の書き込みに対して, 製造メーカは注意を払う必要がある。根拠のない悪意が広まることは, 極めて危険である。このように, 企業にとった良い面と悪い面を含んだシステムであることは明らかであろう。

そのため, 企業によっては, あるいは企業に限らず一般的なオンライン・コミュニティを運営する主体においても, 継続するか, 廃止するかの問題が常に存在する。結論的には, 悪意をもった消費者が存在することを前提にして, コミュニティを継続するか, コミュニティの存在そのものを無視するかの2つになるであろう。

価格設定

インターネットへのアクセスにより, サイトで掲載された情報は瞬時にして, 多くのユーザに公開される。この利点を用いたシステムとして, 同種の製品の企業ごとの価格を公開するサイトが存在する。日本では価格.com があり, 米国では住宅ローンの金融機関ごとの数値を掲載する Proce Watch などがよく知られている。これらは, いわゆる代替性認識効果とよばれており, ある製品には必ず競合他社があり, 消費者はその情報を前提として, 最終的に購入する製品を決定する。

また, このような価格の情報提供とならんで, 一括販売による値引きを積極的に示す商品提示も可能となる。あるいは, 入札の形式を備えたサイトとして運用するなどのケースも存在する。多くの場合, サイトの運営者は, 商品の情報提供と同時に, 在庫確認や配送などの副次的なサービスを請け負っており, これによる収入がサイトの運営費に当てられる。

このように, 一般的には安価な製品を, より大きなバンドルで提供する仕組みが追求されるが, 一方では, サイトでないと購入できない製品をどのように提示するかの追及も行われている。価格だけに注目したサイトでは, いかに安くするかだけが焦点になり, 品質やサービスがともなうかが, いずれ問題となるであろう。

ウェブによる販売チャネル

インターネットを介して行われる商取引, すなわち電子商取引については, 金額ベースでは企業間の取引が大部分を占めている。従って, 電子商取引の将来的な発展は, この B2B の進展に左右されると言える。これに対して, 企業と消費者との間での商取引であるについては, 金額ベースでは相対的に小さいが, その伸び率の大きさから注

目を集めている。日常生活の中でも, インターネット・ショッピングは消費者の購買行動の一部になっており, これに関連した企業も急成長している。

従って, ウェブによる販売チャネルについては, 今後とも, いやがおうにも比較や考察の対象となるであろう。手軽で家にいながら買い物ができることや, 選択の幅が場合によっては店舗からの購入より広くなるケースがある。また, 店舗においてある商品より割安であるなどの利点も有している。

しかし, 一方では, このような傾向は長く続かないのではないかとの悲観的な見方もある。また, インターネット・ショッピングの一回あたりの購入額も, 日本円で1万円以内であり, 市場規模の拡大は望めないのではないかとの意見もある。問題の多くは, 商品の購入だけではなく送料や手数料を含むこと, サイトが貧弱, クレジットカードの不正使用が心配である, 返品できないなどの点にあり, 徐々に解消されてはいるが, 障害となっている。

また, 従来からの企業活動を継続し, この上にインターネット・ショッピングを実現した企業においては, 従来の店舗型の販売とネット販売とを, どのように調整するかも課題である。

3 パーソナライゼーション

3.1 顧客データベース

商品のブランド管理から, これを購入する顧客管理へと移る中で, 顧客データベースをどのように構築し, 利用するかに注目が集まっている。欧米や日本における経験も蓄積されていく過程にあり, その効果と同時に, 問題点も明らかにされつつある。

種々の調査を参考にすると, 顧客データベース導入の効果については, 以下のようにまとめられるであろう (上から意見の上位)。

(1) 反復して購入する顧客の増加
(2) ダイレクトメールによる新規顧客開発の成功率向上
(3) 顧客に合わせた商品販売戦略の作成
(4) 優良顧客の囲い込みの成功

また, 顧客データベースを構築してからの経験年数で見ると, 多くの調査項目において, 導入からの年数が経過している企業ほど, 導入への評価が高いことが検証される事例が多くなっている。この現象は, 多くの情報システム高度化への企業の意見を反映しており, 導入効果が薄いと感じる企業は, 情報システム高度化への取り組みが一時的であり, 単発的に狭い範囲の導入効果とデータ利用に限定していることが検証される。従って, 顧客データベース構築にあたっても, 以下のような情報システム導入と同様なポイントを重視する必要がある。

(1) データ入力だけでなくバックヤードとの連携2次利用を考える
(2) トップを含めて全社的な取り組みとする
(3) 専門の部門を設定する
(4) 継続的な投資を行うが初期設計を重視する

このような, 導入にともなう計画性やその後の活用に関する方法論の違いから, 顧客データベースシステム構築の場合における課題として指摘されるポイントも, 整理されつつある。いくつかの調査事例を参考にすると, 顧客データベース導入における課題として, 以下のようなことがある (上から上位)。

(1) コストや人手がかかる
(2) システムの改編や変更が難しい
(3) データの加工や分析手法が明確でない
(4) 導入効果が明確ではない
(5) データ入力に多大な労力を要する

このような困難性はあるが, 近年では, 大規模なデータベース管理システムを前提としたシステム構築手法の提案や, いわゆるデータマイニング手法の開発により, 数年前よりは, かなり状況は改善されてきていると思われる。

このようにデータの集積や解析ツールにおける技術は改善されると思われるが, 課題となるものは, その活用の方向性であろう。

3.2 データマイニングと顧客管理

顧客データの多くが, 商品の購入時における顧客の記入事項, 会員参加希望への記入事項, あるいは, その他の来店時の記入事項などの他, 懸賞への申し込みなど, 企業と顧客との直接的な接触が契機となっている。これに対して, 現在では, 例えば自治体における住民情報を閲覧した結果を顧客情報とリンクするなどの行為や, 収集目的の異なるデータを他の目的で流用することは, 個人情報保護法に違反する行為であり, 今後は少なくとも社会的に認知されている企業では, 実施されなくなるであろう。

従って, 商品あるいは企業を前面に出し, 顧客がデータを直接的あるいは間接的に企業に提供することが可能となる方法論が必要となる。この一方で, 顧客の目的とする商品選択や検索について, 望んでいることが, かなり異なる事実に対応する必要が強調されている。例えば, 単に関連する分野の商品から 1 つを選択したいのか, やや好みがうるさく, 他の消費者の意見を参考にしたいのかなどの違いである。顧客データベースが初期の顧客開発に役割を果たすとすれば, これ以降のリレーションの維持に役割を果たすものが, 顧客の直接参加によるサイトの運営である。

このような分野の 1 つとして, 顧客ごとに商品選択の方法論を分けて提供することが論じられている。これは実践的に検証されており, 本だけではなく音楽関係など多くの分野で定評のある Amozon.com のサイトでも, この方法論が示されて, 以下のような 4 つの方式により顧客に対する情報の提示が行われる。

(1) 保証提供型

本で言えば, 文学賞を受賞した作品など, 定評のある作品・商品の情報が, 一括して入手できる方法である。これは, しかしながら, 誰でもが労力さえいとわなえれば入手できる情報であり, パーソナライゼーションの範囲ではないとの議論もある。従って, 類型から言えば, 顧客がパーソナライゼーションを拒否しているケースとして分類される。

(2) 協働型フィルター

顧客の好みが主観的な側面が強く, 複雑すぎる場合に適用される方法であり, 顧客の参加が前提となる。まず, ある商品を選択した顧客の属性を, 選択肢への回答という形で記録しておく。次に, 好みがうるさい顧客が到来した場合, 過去に記録した顧客の属性に近いかどうかを識別し, もし, 近い属性が見出されたら, 過去に該当する顧客が購入した商品を, 到来者に提示する。過去の顧客の情報の蓄積と, マッチング処理による商品提示である。

(3)CASE

CASE とは,Computer-Assisted Self-Explication の略語であり, 顧客に対して質問を示し, これに答える形で最良と思われる商品を, 最終的に提示する方法論である。すなわち, 膨大な商品の数を仮定し, 顧客の行動から商品を絞り込む方法である。商品の正確な分類というよりは, 顧客が商品を絞り込む場合の, 選択肢に注目した方法論であるとも言える。

(4) ルール設定型

顧客の母親の誕生日が近づくと, 昨年プレゼントした商品を知らせるなど, いわゆる, ルールに定められた顧客への対応を行い方法論である。この大きなメリットは, 顧客にとっては膨大なアンケートに答える必要がなくなることであり, 企業にとってはパーソナライゼーションの過程を単純化できる点がある。しかしながら, このルール設定型の方法では, 極めて大規模なデータベースにより, 画像情報も含めて大きな顧客情報を管理する必要があること, これらを個別に管理し, オンラインで企業の最新情報のデータベースと結合するなど, 大掛かりなシステムになる課題がある。

以上のような,4 つの顧客のためのツールが存在し, 初期の顧客獲得の段階で, どの方法論が適当であるかを見際める作業が行われる。

4 ブランド形成ステージ管理

4.1 ブランド形成

ブランド形成は,従来のマーケティングにおいても,最終的に到達するための段階として定義されている。ブランド形成については,明確な規定はないが,分かりやす言葉で言えば,商品の評判が安定しており,大多数の消費者が納得する商品であると言えよう。

しかし,すでに述べたように,このブランド形成ののちの段階として,顧客管理が位置づけられており,その意味からも,従来のブランド形成やこの管理とは,やや異なる要素が加わることになるであろう。すなわち,インターネット・マーケティングの視点からの整理が必要である。

ブランド形成,あるいはブランドの維持管理をインターネットの上で実施するためのポイントとして,次のようことがあげられている。

(1) トラフィックの獲得

トラフィックとは,インターネット・トラフィック,すなわちインターネットを飛び交うパケットの頻度を意味するが,この場合には,消費者からホームページへとアクセスされる回数のことを指している。

このトラフィックを生み出す前段となる媒体には,外部のリンク検索,サイトのディレクトリ,広報誌あるいは有料無料広告などがあげられている。サイトを立ち上げて維持管理するには,一定の費用がともなうが,この費用に比べると,サイトが認知されるまでの費用は相対的に大きなものになる。特に,インターネット・サービスが開始された時期とは異なり,現在では極めて多数のサイトが存在しており,消費者がアクセスするに十分であると同時に,他のサイトに埋もれてしまう危険性がある。商品の広告を出す場合に,その商品の名前と同時に,サイトのアドレスが掲載される理由もそこにある。

(2) サイトのドメイン名

しかし,これらのほかにサイトのドメイン名の設定が大きな役割を果たすことが検証されている。すなわち,商品がブランドであると認識される背景には,この商品に関連するサイトがあり,このサイトへ到達するには,簡単なドメイン名アドレスで十分であることが求められる。また,サイトの出来具合が商品の情報へ到達するか,すなわちこのサイトに少しでも長く滞在するかの可能性を大きく左右する。いくつかの統計データが示すように,多くのサイトでは,消費者の見る (ブラウジングする) 時間は極めて小さい。この短時間の間に,消費者に訴えることが求められる。

(3) ポータルでの競合

消費者が企業や商品のサイトへ容易に到達できない場合には,いわゆる,ポータルと呼ばれるサイトにおける検索エンジンを利用することになる。従って,このようなポータルにおける競合において,有利に展開する工夫が必要になる。しかも,検索エンジンは現実に極めて多数のサイトを候補として選択するが,実際にユーザに表示されるものは,この中でも上位に限定されている。相対的に多い場合でも,検索された候補の50%であり,少ない場合は検索された候補の1%未満のもののみが表示される。

従って,サイトの出来不出来を常に検証しておく必要がある。具体的には,複数の検索サイトで,必ずヒットし表示されること,ウェブページのコンテストを用いて上位にあることを確認することである。

(4) ユーザの意見を反映する仕組み

サイトにおける,一方的な企業からの商品の提示だけでは大きな問題があることが指摘されている。サイトが高度に整備され,ここへのトラフィックが多い場合においても,ユーザである消費者の意見を聞けるチャネルを持つ必要がある。事例として,文献においては,インテルが半導体チップの不具合が存在しながら,しかもこれを指摘した関係者の指摘を無視し続けた結果として,この関係者の意見がメールとして,極めて広範囲に広がったことが示されている。従って,ユーザからの重要な指摘を組み入れる仕組みにより,サイトを改善するだけでなく,企業のブランドを高めることができる。このような優良なユーザからの意見を,特別のチャネルで見逃さない工夫もなされている。

4.2 bricks-and-mortal/bricks-and-clicks

インターネットを通じたマーケティング, あるいは電子商取引を実施することが, 必ずしも商品のブランド形成につながらないことが認識されている。いわゆる,bricks-and-mortal/bricks-and-clicks として整理されている課題である。分かりやすい例で言うと, 銀行がインターネットを介したオンラインバンキングをはじめた場合には, 従来の店舗型の方式との競合が発生する。すべての顧客が, インターネットを通じて銀行を利用するのではないので, 両方を維持する必要がでてくる。その結果として, コストがかさみ, 結局は従来型の店舗方式へと回帰することになっている。現在では, インターネット・バンキングは, 電子商取引の決済や, これに関連した業務に限定される傾向にある。

また, 店舗型の販売をしている企業で, 代理店などを通じて販売をしている場合には, 特に事情が複雑となる。従来の店舗販売をしている会社は, 商品の製造元がインターネットというチャネルを通じて販売する方法を同時に選択した場合には, この企業により代理店は軽視されたとの判断をする。その結果として, 店舗での取り扱いを拒否される事態にいたる。

このように, 現在では, インターネット・ショッピングに適合し, 将来の販売増が期待できる商品と, そうではない商品との見極めがなされている。一般的には, 価格の安い商品提供が可能な場合, 希少性や専門性の高い商品が, インターネット・ショッピングに適している。しかし, 紳士用の洋服など試着が必要なものや, 体験しないと良さが分からないものは, 店舗型の販売が有利である。特に, 日本のようなデパートや店舗の充実度の高い国では, 境界が明確である。

5 オンライン・コミュニティ

5.1 コミュニティと企業

オンライン・コミュニティの形式や, これを成功させる要因分析などが継続的になされているが, 以下では, コミュニティのパターンと特徴について整理しておく。コミュニティを, その専門性によりパターン化すると, 極めて専門的なメンバーにより構成されるケースと, 専門性をまったく問わないグループにより構成される場合を両極端として, この中間に多くのバラエティが存在することになる。専門性の強いコミュニティの例として BioMedNet があり, 生物学者や薬学の専門家が参加するものが知られている。企業は, この仮想空間に書店をもうけて広告の掲載料を収入として得たり, あるいは, コミュニティの参加者のメーリングリストを企業へ販売することにより収益を得ている。この場合, 週に約 3000 名の新規参加者に対して, プロモーションのメールを望むかどうかを選択させている。

これに対して, 専門性をまったく問わないものとして RedMole が知られており, 学生に対する職業案内, 大学のカリキュラムの紹介などがなされる。この他に, 授業に関する質問を投稿し, これに対する回答を有償で行うことができたり, 意見を交換するボードが設けられたりしている。また, パートやフル労働の募集案内もなされる。参加者は, 週に約 1 200 の意見を交換している。

これらに共通する点として, 相互の意見交換があること, 特別のグループによる排除がなされないこと, 信頼されるレスポンスが返されることがある。

専門性を基本として形成されるコミュニティについては, 趣味に関連するものも少なくない。コンピュータの機種である Mac の愛好者や, オートバイのハーレーの愛好者のグループがある。しかし, 愛好者の勝手な行動が場合によっては, 企業のイメージを傷つけることもあり (ハーレーの愛好者による無法行為), この場合には, 企業が直接的に良心的な愛好家のグループを形成して, 対抗するなどの手段もとられている。また,Mac を愛好するメンバーによるコミュニティが存在する一方で,Mac を極端に嫌い, 田舎者のコミュニティであると軽蔑するコミュニティ (MacSuck) も存在している。これらは, 極めて排他的な特徴をもっている。

従って, オンライン・コミュニティという, 一見すると自発的に形成されているグループについても, その維持管理や方向性について, 企業は重大な関心を払っている, あるいは払わざるを得ないことに注意する必要がある。商

品のプロモーション, 職業紹介とリクルーティングなど, 企業にとってプラスに作用する側面と同時に, ネガティブに作用する要因が存在する。

企業がコミュニティに関心をもっている事例として, サイトの運営費への援助などの直接的なものをあるが, 参加メンバーの内訳もその傾向を反映している。Week/Harris Poll の実施した 1997 年の調査によると,42%が何らかの形で職業に関連しており,35%が社会的な理由で, また 18%は趣味を愛好する立場から参加している。

5.2 コミュニティ形成の基本視点

企業のとってコミュニティは, 基本的に多くの顧客との関係を構築するために必要とされるが, 企業によっては, 必ずしも積極的なかかわりを行わないケースがある。その代表的なアクションが, コミュニティにおける企業への反論や, 商品へのクレームの掲載について制限を設けることである。よく知られている例として, 英国の食品製造業である Monsanto の姿勢がある。この企業がコミットするコミュニティのサイトへの匿名の投稿は許可されていないほか, 企業のメッセージを前面に出す。これに反して, 石油卸である Shell International Petroleum は, この種の制限を一切行っていない。そのため, コミュニティの議論はいつも活発であり, ロビイストと消費者の双方から注目されている。

しかし, 一般的には, この両極端の間で規制をどのようにするかを探しているのが現状であろう。さまざまな問題を経験した AOL の意見が示すように, 自由な意見や議論は多くの場合企業への批判的見解を許すことになり, 企業はこれを敬遠する。しかし, 意見表明に対する規制を強化することは, 正式のサイトからのメンバーの遊離と, いわゆる正式には認めがたいサイト (unofficial sites) への流出を意味する。

このような場合における企業努力は簡単なものではないが, 解決策としては, これらの悪意あるサイトの意見に対する反論を, 好意的なメンバーが行うことができる, あるいは行ってもらえるような環境を作ることであろう。

6 顧客管理とツール開発

6.1 顧客とクラスタ

以下の各章では, われわれの提案する顧客管理のシステムを述べる。この基本的な方法論は, 既存の管理手法では適応が困難である遺伝的プログラミング (Genetic Programming:GP) による学習を用いており, コンパクトで, 効率的な顧客に関する情報の整理が可能となる。本論文の前半で述べたインターネット・マーケティングを, 実際に実施するためのシステム構築の,1 つの方法論を与える

商品市場における顧客指向の商品開発と, 囲い込み戦略のもとで, 顧客情報を精度良く管理することが必要となっている。情報システムの高度化により, 多量の顧客データを集積することが可能となり, これにともなって顧客の商品嗜好や購買行動における規則性を推定し, 特定の商品を購入する顧客のグループ (クラスタとよぶ) を分類するなど方法論が議論されている。同時に, 人的な作業の限界から, クラスタ分類や検索を自動化し, 効率化することが課題となっている。特に, クラスタ検索においては関連性や外的基準があらかじめ与えられていないデータ集合を検索し, 抽出されたデータ集合の特徴をルールや言語として示す方法が必要となる。

顧客のクラスタ分類・検出に関して, これまでの研究では多変量解析によるクラスタ分析, 多次元尺度法やニューラルネットワーク法が提案されているが, 用いる変数は数値型変数に限定され, 数値的な判別精度が改善されても, これがユーザに分かる言語の形で提供されない問題がある。言語的なルールによる判断が出力されることにより, 分類の根拠が解釈可能な形で示されるメリットがある [4][5]。また,ID3 などの従来の演繹推論の手法では, 数値データとカテゴリデータとが, 同時に 1 つの大きな判別の木構造の形で示されるため, ルールの構造が複雑となる [6]。また, クラスタ検索の手法を開発する場合に, 演繹推論を分類手法として当初の目的から拡張して用いることは, ペアサンプルと呼ばれる対立的 (一方が合格なら他のグループは不合格) であるデータ集合が準備できないなどの理由から困難である。従って, 検索されるべき対象があらかじめ学習データとして与えられてない場合に相当するク

ラスタ検索については, 別の方法を適用する必要がある。

本論文ではGPによるルール推定の手法を用いて, 顧客情報を管理するシステムを提案し, 実際のデータに応用する。GP手法は, これまで関数近似や, エージェントシステムにおける知識表現, 時系列セグメント認識と時系列予測などへと適用され, 有用性が示されている。本論文では, この手法をクラスタ検索へと拡張している。具体的には, 顧客の嗜好を分類し管理する場合に, 顧客の属性を入力として購買するであろう商品を推定する方法をGPによるルール生成により実施する。同時に, 特定の顧客の商品嗜好を推論し抽出するため, データの集合から特徴的なクラスタを検出する手法を提案し, その応用について述べる。GP手法の基本は, カテゴリ化された変数により記述されるデータ集合に対して, 論理演算を実施する木構造 (GPにおける個体に相当する) を多数与えおき, 適合度に応じて遺伝的操作を, 安定的に検出されるクラスタが発見されるまで繰り返し, これにより最終的にクラスタを特徴付ける論理式を求める方法である。そのため, あらかじめ顧客の属性と購入商品とのペアを学習データとして与え, 購入商品をクラスタとした場合に, このクラスタごとの顧客購買行動の推定ルールをGPにより構成する。

GP手法によるクラスタ分類・検索の利点として, 単独のルールだけではなく個体プールとして複数の推論ルールがえられるので顧客の属性の変動に対応して安定的な分類が可能となることがある。

応用例として, 人工的に生成した顧客購買行動に対する本論文の手法を適用するとともに, 実際に観測される商品購入行動を推定するシミュレーションを実施し, その有効性を確認する。

6.2 GPによるクラスタ分類・検出システム

まず, 最初に本論文において述べるGPによるルール推定を基礎としたクラスタ分類と, 顧客情報管理システムの関連について述べておく。なお, 以下では特に断らない限り, ある特定の特徴をもとにデータ集合から抽出されたデータのグループをクラスタ (cluster) とよぶことにする。

顧客情報を集積することにより, 顧客の嗜好傾向を推定したり, 商品開発に反映させることは, 多品種少量生産や短い商品のライフサイクルのもとでは, 極めて重要な要因となっている。顧客情報管理システムとして, よく利用されているものに, 関係データベースや, これを一般化したオブジェクトデータベースがある。関係データベースにおいては, 検索コマンド (クエリー:query) を作成し, あるクエリーに適合するレコードを抽出することに重点が置かれている。しかし, この方法では, 最初から検索するクエリーを人的に作成する必要があること, 従って抽出されたクラスタの特徴が分かっているので, 顧客を検索する以外に利用価値がない問題がある。

一方では, これとは逆の方向である推論ルールを求める方法が議論されており, その一部はデータマイニング手法として開発されている。簡単な事例として, 特定の商品を買う顧客は, 同時に他のどの商品を購入しているかを推定する方法である。あるいは, 特定の商品を購入する顧客の特徴 (プロフィール) を, 言語的に示す方法である。後者の商品選択の事例とは直接の関係はないが, 消費者ローンで, 審査に合格しなかった顧客の特徴を言語的に出力するシステムも存在する。これらの方法は, 特定の商品を購入した顧客をクラスタとしてとらえ, このクラスタの特徴を言語的に出力する方法であると言えるであろう。もちろん, このような顧客を特徴ずける方法としては, 従来手法である多変量解析においてもクラスタ分析や多次元尺度法, あるいはこれらの結果をこの視覚化などの方法が利用可能であるが, 数値的な説明だけが可能であるなどの問題がある。

以上のようなことを考慮すると, 顧客情報管理システムの備えるべき機能とし, クエリーによる顧客の検索と同時に, 特定の商品を購入した顧客の特徴づけを行い, しかも, この結果を言語的に出力すること, あるいは, ある基準で選択したクラスタに共通する性質を言語的に見出す機能が求められていると言える。

なお, 言語的にクラスタ分類の結果を出力する方法として, 演繹推論などの手法があるが, 本論文で示すGPによる手法は, その分類精度において従来手法より優れており, また多様な顧客の属性に対応できる利点がある。

6.3 GPによるクラスタ分類システムの構成

以下では, 最初に, 本論文で述べるGP手法によるルール推定を基礎とした顧客情報管理システムについて, 第1番目の構成要素であるクラスタ分類システムの構成について示す [9]。システムの概要を図1に示す。まず, システム全体で用いるデータ集合については, 次のようにまとめられる。データは1件ずつレコードの形で格納されており, レコードの項目 (フィールド) は数値データ, あるいは質的データ (カテゴリカルデータ) であると仮定する。なお, 本論文では, レコードの分類は集合としてなされるので, その所属をクラスタとよぶ。一方, カテゴリという言葉は, フィールドに格納されるデータが, いわゆるカテゴリカルデータであるかどうかを示すときに使うことにする。

なお, 数値データはそのまま用いることもできるが本論文のシステムではカテゴリ化されると仮定する。従って, 数値データはいくつかのグループに集約して,1つのカテゴリに変換して用いることにしている。その方法論ついては, あとで議論する。これらを含めてすべてのカテゴリカルデータは, クラスタ分類システムを構成するための, 学習データとして準備されていると仮定する。具体的には, 学習データとして用いる顧客については, 例えばどの商品を購入したかなどのクラスタが, すでに外的基準として与えられていると仮定する。

カテゴリとカテゴリ値・カテゴリ変数

以下では, 顧客に関するカテゴリデータがデータ集合して格納されていると仮定し, カテゴリに割り当てられた値をカテゴリ値とよぶ。カテゴリの値を代入する変数をカテゴリ変数とよぶ。このクラスタ分類システムの目的は, 特定の商品を購入した顧客に関するカテゴリカルデータを用いて, どのような推論結果がルールとして得られるかを求めることにある。カテゴリカルデータは, いわば論理値として処理できるので, 本論文では, 最終的なクラスタ分類を値として出力するシステムを, プロダクションルールにより生成する。例を次に示す。

```
if  v1=1 or v1=2 and v8=1 then class="A"
```

この例では, 第1番目のカテゴリが1であるか, もしくは2であり, 第8カテゴリ変数の値が1であるならば, 購入商品は A であることを推定値として出力するルールとなっている。

このようなルールは木構造で示され, 実際には等価な前置表現によりGP手法における個体として格納される。ルールを記述する木構造である論理式の形はさまざまなものが可能であるが, 以下では, 比較的簡単な表現を用いている。詳細は後述するが, カテゴリ k において顧客のとるカテゴリ値が j であるとき論理変数 (カテゴリ変数)X_{kj} は真の値をとり, これ以外は偽となる論理変数としておく。プロダクションルールは, これらの論理変数を用いた論理式で記述される。

学習データ

準備された学習データを用いて,GP手法における個体のプールを構成しその性能を改善していく。この場合, 顧客に与える購買商品のクラスタごとにプールを構成する。すなわち, 購入商品ごとに顧客データが準備されており, カテゴリの記号として表現されていると仮定する。図1上部に示す学習フェーズ (Learning Phase) では, それぞれの購入商品 $R_i(i=1\sim n)$ ごとにこのような学習データが準備されていることを意味している。すなわち, 外的基準として購入商品が分かっていると仮定し, この購入決定に相当する顧客の属性を示すカテゴリカルデータを学習データとして準備しておく。

個体の生成

購入推定のルールは論理式で記述されると仮定し, 論理式は論理変数を終端記号とし, 中間の節に論理演算記号を配置した木構造で表現される。ただし, 一般的な形を許すとアルゴリズムが複雑となるので, 本論文では木構造を2分岐の形式に限定する。このような制限のもとでは関数近似の算術式において

変数 → 論理変数

演算子 → 論理演算子

のような置き換えを行うと, 論理式の個体表現を求めることができる。

個体の初期値をランダムに生成しておき, それぞれの個体の推定能力を計算する。学習データを用いて, 購入商品 I の推定値 $\hat{I}$ を個体ごとに計算し, この計算値があらかじめ与えている購入商品 I にできるだけ一致する方向に,GPによる学習を進める。具体的には, ある個体により顧客属性から購入商品を推定し, その推定値が実際の購

入データと一致する確率を求める。この確率が大きいほど, この個体によるクラスタ推定が良好であることを意味するので, 後段の遺伝的操作において, この個体がより個体プールに存続する確率が高まる。

購入商品の推定

次に, 顧客のデータが存在して, その購入希望商品が不明であるケースについて, 予測を図1下部に示す推定フェーズ (Prediction Phase) において実施する。そのため, 学習フェーズで求めた個体プールのすべてに, このルール当てはめを実施する。当てはめの予測結果を, 適合度の大きな個体から数個を選択し, 適当とされる予測値をきめる。通常の適合度を用いた判別では最高の適合度をもつ個体の判別結果だけで推定できる。しかし,GP における固体の特性を生かすため, 適合度の高い個体から得られる推定値から, 相対的に頻度の高い推定値を最終的な決定としている。

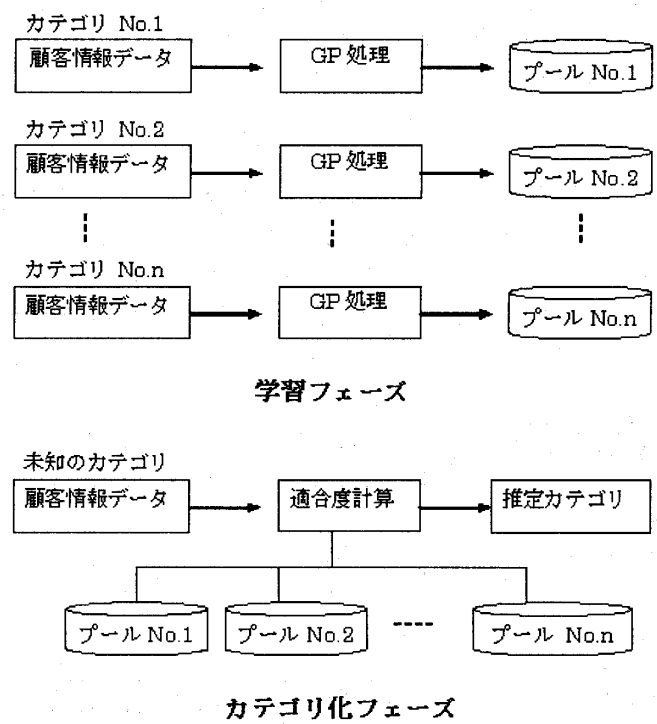

図 1: 顧客クラスタ分類システムの構成概要

6.4 数値データを用いたカテゴリ化の概要

本論文では, 顧客に関するデータはすべてカテゴリデータであると仮定している。しかし一般には顧客に関するデータには数値データも存在する。この数値データはそのまま顧客のクラスタを分類したり検索する場合に直接用いることも可能であるが問題がある。まず, 数値データは連続データであるため無数のルールが必要になることがあり, また, 最終的には, この数値を解釈する必要がある。このような理由により, 本論文のシステムでは, 数値データは単独で, あるいはグループとしてまとめて1つのカテゴリ変数に変換することを行っている。この変換の方法は, さまざまな実現方法が可能であるが, システム構成の統一性を保つためと, 推定の精度を維持するため,GP 手法を用いることとする。

なお, このカテゴリ化のシステムの構成は, クラスタ分類推定のシステムと同様の考え方で構成できるので, 以下では簡単に要点だけを説明する。

学習データ

システムの概要を図2に示している。システムは, 数値データを入力データとして, この数値から判断される中間的な分類 (カテゴリ) を推定し, 出力するシステムとして構成される。数値データから推定される中間的なカテゴリは複数存在すると仮定し, これらのそれぞれに,1 つの木構造の近似関数の集合 (これを GP における呼び方にならって, 個体プールと呼ぶ) が対応している。図2上部に示す学習フェーズ (Learning Phase) では, それぞれのカテゴリ $i(i = 1 \sim n)$ ごとにこのような学習データが準備されていることを意味している。すなわち, 外的基準としてカテゴリが分かっていると仮定し, このカテゴリに相当する数値データを学習データとして準備しておく。

個体プールの生成

最初に, 学習データを用いて分類すべきカテゴリごとに推定する関数の近似形 (GP における個体に対応する) を求める。この近似に GP 手法を用いる。多変量解析における判別関数の構成と同様に, 数値データの関数を仮定した場合に, それぞれのカテゴリに属する数値変数のデータ $x = (x_1, x_2, ..., x_m)$ を与えた場合にだけ関数 $f(x)$ が大きくなるように関数 $f(x)$ の近似を行っていく。個体の中で関数の数値が大きなものだけが個体プールに残り, 更に, これを用いて関数近似を改善する方法を用いている。これを示したのが図2上部に示す学習フェーズ (Learning Phase) である。

カテゴリの推定

以上のような学習を適用して, 個体プールを準備しておく。次に, カテゴリが未知である数値変数のデータ $x = (x_1, x_2, ..., x_m)$ を入力した場合に, このカテゴリを決定する必要がある。これを示したのが図2下部のカテゴリ化フェーズ (Categorization Phase) である。この場合, すべてのカテゴリ k の個体プールの個体 j である関数 $f_k^j(x)$ に対して数値データ x を入力変数として代入し, その関数値が最大となる個体がプール k に属している場合, この数値データのカテゴリを k であると推定する。

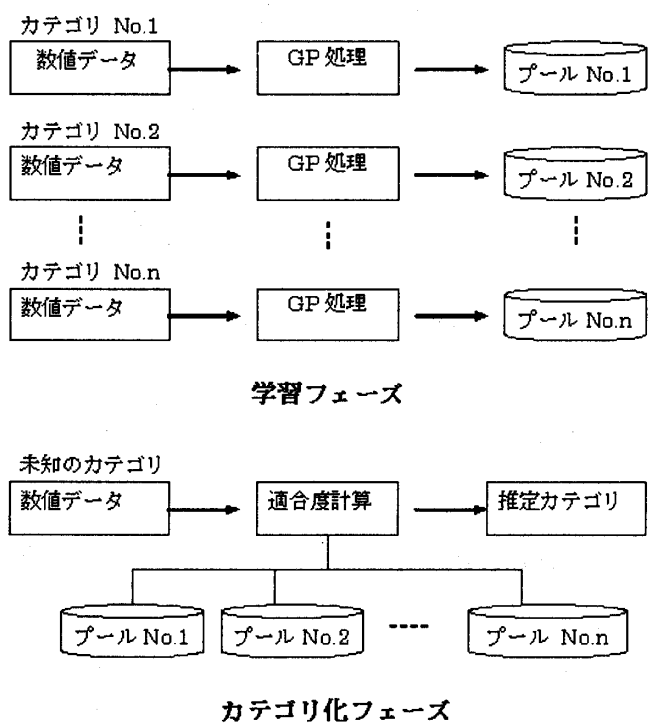

図 2: 数値データのカテゴリ化システム構成の概要

6.5 GP手法の基本と顧客クラスタ推定

本論文で用いるGP手法については,すでに多くの記述があるので,以下では要点のみを簡単にまとめておく[10]-[20]。なお,以下では分かりやすくするため,GPによる関数近似をとりあげるが,簡単な置き換えによりクラスタ分類・推定ルールの自動生成へと拡張することができる。

GPはGA(Genetic Algorithm:遺伝的アルゴリズム)の1つの拡張であり,個体はGAのようにビット列ではなく,数学演算のための演算子,変数からなっている[10]-[20]。GPのシステムは基本的に,3つの部分からなりたっており,その第1番目は個体の表現である。方程式は木構造で表現できるが,これを前置表現(prefix representation)により置き換えておく。例えば,方程式$f(x) = [3x_1 - x_2] \times [x_3 - 4]$は,次のように表現する。

$$\times - \times 3 x_1 x_2 - x_3 4 \tag{1}$$

このような前置表現のそれぞれを個体とよぶ。

次に必要なのが,GPにより表現された個体の解釈である。これには式(1)に示すような式により表現された方程式の右辺の形をもとにして,関数の値を求めることである。個体により計算された関数の値$\tilde{f}(x)$と観測された時系列データ$f(x)$との2乗誤差は近似度を与えるので,この逆数により個体の適合度を定義する。

第3番目に必要なのが,GPにより個体を性能の良いもの(今の場合には関数近似の能力の高いもの)に変換していく方法である。個体の集合(プール)の能力をたかめることは,個体に対して交差処理,突然変異処理を行うことにより可能である。このため,*StackCount*というカウンタを用いる。

*StackCount*の値は,前置表現で表現された個体のストリングを左側からサーチしていき演算記号に出会うとその数値を1つ増やし,被演算子に出会うとその数値を1つ減らす操作を実施した結果である。個体のストリングの全体をサーチし終えたあとに,*StackCount*の数値は必ず1になる。従って,GPにおける初期個体を生成するときに,この条件を満足しないものは個体として採用しない。適合度に応じて2つの個体A, Bを選択したあと個体Aの交差位置を乱数により選択し,この位置における*StackCount*を計算しておく。次に,個体Bについて同じ*StackCount*をもつ位置を検出し,等確率で,ある1つの交差点を確定する。これらの位置を境界として,それぞれの個体の前半と後半を,相互に交換した個体がoffspringとして生成される。

以上のような方法をアルゴリズムとしてまとめると,次のようになる[10]-[20]。

(ステップ1)

乱数を用いて被演算子,演算記号の並びからなる初期個体のプールを構成する。個体の表現の妥当性を,すでに述べた*StackCount*を検査することにより行う。

(ステップ2)

個体に表現された関数をもとに,それぞれの個体により得られる予測値を求める。これをもとにして,個体における適合度を求める。

(ステップ3)

次に示す適合度から変換された確率に応じて,2つの個体iが選択される。

$$p_i = (S_i - S_{min}) / \sum_{i=1}^{N} (S_i - S_{min}) \tag{2}$$

ここで,S_{min}は適合度の最低値,Nはプールの大きさである。この2つの個体に対して遺伝的操作を行い,生成された新しい個体を次のステップにおける代替個体のプールであるP-Bに格納しておく。このような新しい個体の生成を,規定回数繰り返す。新規個体の生成が終了したら,プールP-Aの個体の中で,相対的に適合度の低い個体を,プールP-Bの個体により置き換える。

(ステップ4)

ステップ2からステップ4までの交差処理を,決められた個数の個体に適用し,新しい個体のプールを作成したあとに,次に示す突然変異を実施する。

G-突然変異:グローバルな突然変異を意味し,2つの個体に対する交差処理である。ただし,今の場合には,個体Aは

選択され突然変異を適用する個体であるが, 個体Bは, 初期個体の発生と同じ手順を用いて一時的に発生させた作業用の個体である。交差点を適切に決めたのちに, 個体Aの後半を個体Bの後半と交換する (個体Bに対しては, 特に何も操作や保存はしない)。

L-突然変異:任意に個体を選択して, この個体の被演算子, 演算記号の部分を, 任意に選択した被演算子, 演算記号により置き換える。

(ステップ5)

ステップ2からステップ4までの操作を規定回数繰り返す。

6.6 GPによる第2階層論理式への遺伝的操作

これまでの研究においても,GP手法を論理式への遺伝的操作に適用し, プロダクションルールの改善をはかる方法論へ用いている [14]-[16]。そのためには, 論理式の形式を2項演算の形で2つの命題を論理演算子で結合した場合に限定する必要があるが, 大きな制約ではない。従って, 基本的には論理式のレベルにおけるGPによる遺伝的操作は, 算術式における遺伝的操作と同様に実施できる。具体的には, 次のような置き換えを行う。

数値型入力変数 vi → 論理変数 Xi

算術演算子 $+, \times$ → 論理演算子 And,Or

本論文のシステムでは, 第2階層において, カテゴリ識別を記号として与えた系列を入力として, 格付を予測するルールをGPにより学習・改善していく方法を用いる。これにはさまざまな方法が考えられるが, 以下では, 次のような比較的簡単な方法で, 識別ルールを記述する論理式を表現する。

いま, 財務カテゴリを含めて複数のカテゴリ $k, k = 1, 2, ..., K$ について, 企業の特性が与えられていると仮定する。これらのカテゴリに対応する変数を $v_1, v_2, ..., v_K$ としておき, カテゴリのとる値を, 簡単のためと 1,2.3,... しておく。例えばカテゴリ 1,2 の値が 1,3 である場合には,$v_1 = 1, v_2 = 3$ のようになる。従って, これらを論理命題として結合したものを, クラスタ分類推定のルールを生成するプロダクションルールとして記述することができる。クラスタの予測として,A,B など比較的簡単な分類だけを取り上げる。例を次に示す。

if $v_1 = 1$ and $v_2 = 3$ or $v_3 = 1$ and $v_1 = 3$ then A

更に簡単化を行うと, 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。

$$X_{kj} = \begin{cases} True, & \text{if } v_k = j; \\ False, & \text{otherwise} \end{cases} \tag{3}$$

それぞれの個体は, このような論理変数を含んだ論理式を推定のルールとして表現している。論理式表現における個体の適合度は, 次のようにして計算できる。

(1) 論理値の計算それぞれのカテゴリ t_i において記号 1,2,... が出現しているかを検査し, 論理変数 X_{ij} の値を求める。

(2) 論理式の解釈

命題は1つの論理値として与えられるので, 論理演算子を考慮しながら, これを含む論理式の値を計算する。

(3) 適合度の計算

学習に用いている顧客ののカテゴリ記号列のデータに対して, 上のような論理式を用いたクラスタを推定する。この推定結果と, 観測されたクラスタ (実際に与えられた購入商品) とを比較する。推定と実際に発生したデータが同じなら, この個体の適合度を増加させる。

7 顧客クラスタ推定の応用例

7.1 人工的データによるシミュレーション

以下では, 実際に存在する顧客データを用いてクラスタ分類の推定を実施する前に, 本論文のシステムによってどの程度の顧客クラスタ数の分類ができるかの見通しをたてるため, 人工的なデータを用いて検討する。最初に, 顧客の属性がカテゴリ変数として記述されており, そのカテゴリにあらかじめ定められた確率でノイズが含まれているケースを考え, シミュレーションにより分類可能性を調べる。

顧客の選択する商品をクラスタ $L_1, L_2, ..., L_K$ として表し, クラスタの数 K は $K = 3, 6, 9$ の 3 つの場合を考察する。第 1 階層の数値データのカテゴリ化の性能は, 次に検討するとし, 以下では, 顧客の属性はすべてカテゴリ変数だけで記述されていると仮定する。シミュレーションでは, 次のような比較的簡単なケースを仮定する。顧客の属性を記述するカテゴリ変数を $v_1, v_2, ..., v_m$ としておき, クラスタ $L_1, L_2, ..., L_C$ の順にカテゴリ変数の値 $v_1, v_2, ..., v_m$ が, $1, 2, ..., K$ となる値をとるとする。すなわち, クラスタ L_1 の顧客のカテゴリ変数は基本的にはすべて 1 となると仮定する。同様にクラスタ L_2 のカテゴリ変数は, 基本的にはすべて値 2 をとると仮定する。

このままのカテゴリ変数の与え方のもとでは, 全てのクラスタはカテゴリ変数を入力とする判別システムで,100 %正しく判別される。そこで, 以下のように, カテゴリ変数にノイズを導入する。クラスタ L_1 の顧客のカテゴリ変数 $v_1, v_2, ..., v_m$ は, すべてが 1 ではなく, ある確率 p(以下のシミュレーションでは $p = 0.3$ としておく) で,1 以外の値に変更される。同様に, クラスタ L_2 の顧客のカテゴリ変数の値を, 確率 p で 2 以外の値に変更する。

シミュレーションの条件を, 次のようにする。

格付ランク数:3,6 および 9
カテゴリ変数 v_i の数:6
カテゴリ変数のカテゴリ数:3

GP 適用の条件は, 以下のようである。

個体の長さ:20
個体に含まれる演算子:And, Or
個体に含まれる変数:式 (3) に示す論理変数 X_{ij}
個体プールの中の個体数:1000
GP 適用回数:600

表 1 には, この場合のシミュレーション結果を示している。表 1 では, クラスタの区分 (これをランクとよんでいる) が $R_1, R_2, .., R_9$ として表示され, この総数が, それぞれ 3,6,9 の場合のシミュレーション結果を示している。表では, あらかじめ与えたクラスタの値ランクと, ルールにより推定されたランクとが一致する割合を示している。

このような条件のもとで, 本論文のシステムを用いてクラスタが正しく推定できる確率をシミュレーションにより求める。これを示したものが表 3 である。この表から分かるように, クラスタの数 C が 6 以下である場合には, クラスタが正しく推定される確率は 80% 程度であり, 実際に応用する場合にも大きな支障はないと考えられる。これは, 人工的なデータ生成ではあるが, 一定の方法で GP 個体プールを生成しているので, データサンプルのばらつきを吸収している効果が見られるためと予想される。しかしクラスタが 9 の場合には正しいクラスタの推定精度は極端に低下し, このままでは実際に応用するには問題がある。クラスタ推定の精度が低下する理由は 1 つのクラスタに対する学習サンプルとこのクラスタ以外の学習サンプル数との割合が 1/8 という小さな数となり望ましい学習が達成できないことにある。

以上のようなことから, 本論文の顧客カテゴリ分類システムで分類可能なクラスタ数の最大値は,6 程度であるといえる。これを, 直接, 実際の顧客クラスタ分類の問題に適用することはできないが,1 つの目安を与えている。すなわち人的な方法により実施されている詳細化, つまりクラスタ分類のプロセスを相対的に多くの商品数のクラスタ分類にまで拡張し, 本論文の手法により再現することは難しいことが分かる。

しかしながら, 顧客属性を 1 つに集約するのではなく 2 つ以上のグループに分け, 詳細なクラスタ分析を分解す

ることは可能である。例えば, 主要な方法として, 顧客属性の中で財政的なデータから得られる側面のクラスタと, 生活環境からみた側面からみた格付を, 行列の行 (横方向) と列 (縦方向) に配置し, この行列の縦横のクロスする部分に, 細かなクラスタを更に配置することが行われている。この場合の財務的側面と生活側面のどちらもクラスタは5程度である。以上のことを考慮すると, 基礎的なクラスタのランクを6程度にすることは適切であり, この意味で本論文の手法は有効であると言える。

表1. 人工的データに対するクラスタ推定の結果 (p:%)

区分	L1	L2	L3	L4	L5	L6	L7	L8	L9
3ランク	88.3	84.9	90.0	-	-	-	-	-	-
6ランク	86.0	82.7	75.9	81.3	79,0	82.9	-	-	-
9ランク	43.5	56.4	36.3	68.3	46.7	32.1	43.3	32.1	55.7

7.2 数値データからのカテゴリ推定

次に, 数値データをまとめて1つのカテゴリ変数の値に変換する第1階層のシステムの性能について, 人工的なデータに対するシミュレーションにより明らかにする。以下で示す性能評価の方法では, 基本的に, やや理想化した分かりやすいケースをとりあげている。

顧客に関する数値型変数 $x_1, x_2, ..., x_n$ が与えられており, この変数を入力とする変換システムをGP手法により構成し, カテゴリ y の値を求める問題を考察する。カテゴリ変数 y の取りうる値 (値域) を,$1, 2, 3, ...K$ としておく。数値型変数 x_i の分布は, 平均が μ_i, 分散が σ^2 である正規分布を仮定する。変数の分布と推定されるカテゴリ値の精度との関係を求めるため, 以下のような簡単化をはかる。変数の確率分布のパーセンタイル点を計算し, これを順に $Q_1, Q_2, .., Q_K$ と呼んでおく。学習データおよび検証データにおいて, 顧客の属性として与えられる数値変数のカテゴリ化の値が k である場合には, この数値型変数の値は k パーセンタイル点である Q_k の値をとると仮定する。すなわち, 第2階層のクラスタ分類推定の場合と同様に, 例えば, カテゴリ化の結果である外的基準が1である場合には, すべての数値型変数は第1番目のパーセンタイル点である Q_1 をとると仮定する。

しかし, この前提ではすべてのカテゴリ化が正しく行われることが保証されているので, 以下に示すようなノイズを導入する。カテゴリ1となるべきデータの数値変数 $x_1, x_2, ..., x_n$ は, すべてが Q_1 ではなく, ある確率 p(以下のシミュレーションでは $p = 0.3$ としておく) で,$Q_1$1 以外の値に変更される。同様に, カテゴリ2となるべき外的基準の数値変数の値を, 確率 p で Q_2 以外の値に変更する。

このような条件のもとで, 本論文で示す第1階層のシステムにカテゴリ値が正しく推定できるかを求める。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。

個体の長さ:20
個体に含まれる演算子:$+, -, \times, abs$
個体に含まれる変数:24 指標全部および2グループに分離した場合
個体プールの中の個体数:1000
GP 適用回数:600

これを示したものが表3である。この表から分かるように, クラスタの数 C が6以下である場合には, クラスタが正しく推定される確率は80% 程度であり, 実際に応用する場合にも大きな支障はないと考えられる。しかしクラスタが9の場合には正しいクラスタの推定精度は極端に低下し, このままでは実際に応用するには問題がある。クラスタ推定の精度が低下する理由は1つのクラスタに対する学習サンプルとこのクラスタ以外の学習サンプル数との割合が1/8 という小さな数となり望ましい学習が達成できないことにある。

表2. カテゴリの認識結果 (単位:%)

カテゴリ	1	2	3
認識結果	86	84	87

表 3. 財務カテゴリの認識結果 (2 種類の指標を別々に使用, 単位:%)

	収益性指標			キャッシュフロー指標		
財務カテゴリ	1	2	3	1	2	3
認識結果	81	72	66	73	63	76

7.3 実際のデータを用いたシミュレーション

以下では, 実際に観測される顧客情報を用いて, 本論文で示すシステムによる推定の性能を評価する。用いる顧客情報は、研究名目のために企業より提供を受けた POS データを元にして作成されており、「性別」「年齢」「職業」「未既婚」「年収」「居住地」「家族構成」「住居形態」「購入ブランド」の属性が 5 段階のカテゴリデータ (名義変数) として与えられている (表 7)。数値データからの離散化においては, 基本的にデータの統計的な分布を求め, そのパーセンタイル点を参考点として, 離散化を実施している。このうち性別～住居形態まで 8 属性のカテゴリデータを元にして、GP により購入ブランドのカテゴリデータを算出するルールを作成する。

ブランドカテゴリ数:4
カテゴリ変数の数:8
カテゴリ変数のカテゴリ数:5

GP 適用の条件は, 以下のようである。

個体の長さ:20
個体に含まれる演算子:And , Or
個体に含まれる変数:式 (3) に示す論理変数 X_{ij}
個体プールの中の個体数:1000
GP 適用回数:1000

GP によるカテゴリ選別ルールを用いた分類との性能比較のため、ベクトル空間で表現されたデータに対するカテゴリ化アルゴリズムとして一般的な「余弦法」を合わせて行う。まず各カテゴリに所属する顧客情報を各属性についてのベクトル表現と捉え、それらから計算される平均ベクトルをカテゴリベクトルとして考える。これと未分類の顧客情報との余弦を計算することで、その値が 1 に近いほどそのカテゴリへの親和性が高いとするものである。

表 4. 用いたカテゴリカルデータとその計算手法

カテゴリ変数	計算方法
v1:性別カテゴリ	顧客の性別をカテゴリ化したもの
v2:年齢カテゴリ	顧客の年齢を年代ごとにカテゴリ化したもの
v3:職業カテゴリ	顧客の職業をカテゴリ化したもの
v4:未既婚カテゴリ	顧客の未婚・既婚をカテゴリ化したもの
v5:年収カテゴリ	顧客の年収をカテゴリ化したもの
v6:居住地カテゴリ	顧客の居住エリアをカテゴリ化したもの
v7:家族構成カテゴリ	顧客の家族人数をカテゴリ化したもの
v8 住居状態カテゴリ	顧客の住居をカテゴリ化したもの
v9 ブランドカテゴリ	顧客が購入したブランドをカテゴリ化したもの

表5. 購入ブランド推定の結果

購入品目	余弦法	GP ルール
物品 A	40%	55%
物品 B	51%	60%
物品 C	55%	62%
物品 D	43%	61%

表5の結果より,GP 手法を用いたことで従来の余弦法よりも高いカテゴリ推定能力を有することが分かる。

8 GPによる顧客クラスタ特徴の記述

8.1 顧客クラスタ特徴記述の必要性

近年, 情報システムにおける装置の大容量化にともない多量の蓄積データが利用可能となっており, これらのデータを分類・検索した結果を, さまざまな意思決定に用いることが重要となっている [1][2]。特に, 一定の基準で抽出・分割されたデータの集合 (これを, 以下ではクラスタとよぶ) を特徴ずける手段を明らかにすることで, より高度な情報を提供することが可能となる。

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。クラスタ分類に関して, 従来より多変量解析法などを基本としたクラスタリング手法が知られている。この方法をクラスタ特徴記述に拡張することも可能であるが, しかしこの手法は数値的な結論をベースにしており, クラスタの特徴について言語的に説明できない。言語的な記述を利用する方法として, ニューラルネットワーク構成を簡素化する方法や, ID3 などの従来の分類手法を拡張して用いることも考えられるが, 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 [3]-[6]。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。

本論文では,GP 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を提案し, その応用について述べる [7][8]。具体的には, カテゴリ化された変数により記述されるサンプルに対して, 論理演算を実施する木構造 (GP における個体に相当する) を多数与えおき, サンプルに対して論理式が成立する割合として定義する個体の適合度に応じて, 安定的にクラスタに対してだけ論理式が成立するまで個体に対する遺伝的操作を繰り返し, これにより最終的にクラスタを特徴付ける論理式を求める方法である。

GP 手法は, これまでカオス力学系における関数近似や, エージェントシステムにおける知識表現, 時系列セグメント認識と時系列分類・予測などへと適用され, 有用性が示されている [9]-[18]。本論文では, クラスタ内のサンプルだけが抽出される (ヒットする) 方向に, 検索ルールを GP 手法により改善していく。

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。

応用例として, 人工的に与えたクラスタを用いた性能評価と, 個人へのローン審査データを用いた事例について述べ, 本システムによるクラスタ特徴記述が良好であることを示す。また, これらの他に 8 種類のデータ集合に対する適用結果を示す。

8.2 GPによるクラスタ特徴記述システム構成の概要

本論文で述べるGPによるクラスタ特徴記述システムシステム概要は,以下のようになる[7]。図3には,システム構成の概要を図示している。

(1) サンプルをカテゴリ変数で記述する

サンプルを記述する変数には,数値型変数とカテゴリ型変数が存在するが,数値型変数についてはカテゴリ化サブシステムにより,1つのカテゴリ型変数の集約されると仮定する。数値型変数をカテゴリ化する方法については,さまざまな手法が適用可能であるが,われわれが以前示したGP手法による方法も適用可能である[16][17]。しかし,このような適用手法は,以降の議論と大きな関連性はないので,詳細は省略する。このような前処理により,それぞれのサンプルは,カテゴリ変数だけで記述することができる。

(2) クラスタ記述の論理式個体の初期値生成

カテゴリ変数に対する論理変数を変数とする論理式(詳細は後述する)により,クラスタを特徴付けるルールを表現する。論理式を表現する個体は,システム構成を簡単化するため2分岐構造に限定しておく。木構造をなす論理式が正当なものであるかを検査する方法を用いながら,ランダムに複数個(例えば1000個)を初期値として生成する。これらは,GP手法における個体とよばれ,個体の集合をプールとよぶ。

(3) 個体の適合度の定義とGP適用

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。

次のような指標を定義する。

$$y_k = T - h_k^2/n_k \tag{4}$$

式(1)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の倫理式が真となる数

h_k : クラスタcに含まれるサンプルで個体kの倫理式が真となる数

T : クラスタcのサンプル数

個体kの適合度f_kは,式(4)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。

$$f_k = (a + y_k)^{-1} \tag{5}$$

式(5)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(5)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。

(4) クラスタ特徴記述の終了

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。

8.3 クラスタへの分解

本論文では,あらかじめクラスタが与えられた場合に,その特徴を記述するルールをGPにより推定することに重点が置かれている。従って,クラスタをどのように抽出するかについての詳しい議論は行わない。しかしながら本論文で提案するシステムの性能を評価するシミュレーションを実施する場合に次のような点に留意している。

(1) クラスタ抽出のための数値型・カテゴリ変数を限定しない

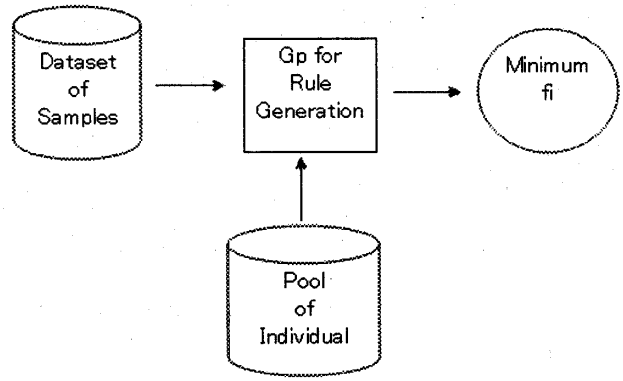

図 3: クラスタ特徴記述システム構成の概要

クラスタ抽出を行う場合に, 少数の変数にだけ注目するのは問題がある。例えば,1 つのカテゴリ変数だけに注目すれば, これを基準とするクラスタが形成されるのは自明である。従って, 従来のクラスタリング手法を基本として, 本システムを検証するクラスタを抽出するが, 用いる数値型変数およびカテゴリ変数の個数を, 最低でも 6 個以上としている。

(2) クラスタの安定的抽出

クラスタリングに用いる変数の組み合わせにより, 極端にクラスタを構成するサンプル数に差が発生しないようにしている。極めて稀にしか発生しない事象を特徴付けることも重要であるが, 本論文では, クラスタ特徴の記述能力の基本的な性能を評価している。

(3) クラスタ抽出手法の統一化

クラスタを抽出する方法論には, クラスタの代表値の定義や, サンプルとこの代表値との距離を定義する問題が含まれ, 場合によってはクラスタの様相が大きく変動することがある。しかし, 本論文では, このようなクラスタの抽出そのものを問題としてはいないので, いわゆる重心法とよばれる, ごく一般的な手法を用いている。これ以外のクラスタリング手法による差異は, 今後の検討課題としている。

なお, 本論文の手法の有効性を確認したシミュレーション実験では, 以上のような点に留意しておけば, クラスタ抽出の方法の違いにより, カテゴリ変数をもとにして記述されたクラスタ特徴記述のルールに, やや変化が見られることが, 極めて微小な範囲にとどまっていることが確認できる。従って, 以下では, すでにクラスタは分離抽出されており, そのクラスタの特徴記述だけが課題であると仮定する。

8.4 German Creit 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, 本論文の手法の適用可能性を議論する。このデータはドイツの消費者ローン会社で実施された 1000 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 [3][22]。

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, 本論文で示す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 1000 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。

クレジット期間, クレジット額, クレジット利率
現住所での居住期間, 年齢, 当会社銀行でのクレジット開設数
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, 以下のような13個のカテゴリ変数となる。

x1:手形口座開設の内容 (4 カテゴリ),x2:契約継続月数 (数値) x3:クレジット履歴 (5 カテゴリ),x4:借入目的 (11 カテゴリ) x5:預金口座内容 (5 カテゴリ),x6:保証人の有無 (3 カテゴリ)

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:1000

表6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 (5) に示す h, n, y が最適となる個体の値を, 主要なGP世代ごとに示している。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4に示している。この表4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。

表6 h, n, y, N_{GP} の間の関係例

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

表7. 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}

9 GPによる文書分類・検索システムの構成

9.1 顧客管理と文書分類・検索システム

インターネット・マーケティングの環境においては, 顧客からの意見の集約や, さまざまに収集された文書を, 管理することが必要になる。このような意味で, 以下では,GPによる文書分類・検索システムの構成について述べる。

情報システムの高度化にともない, 多量の文書データを格納し, 参照することが可能となり, さまざまな意思決定の基礎データとして利用されている。また, インターネットを通じて関連する文書を探索する方法も一般化している。このような文書検索の場合には, 従来のキーワードやアブストラクトの情報を用いた検索だけではなく, 記述されている内容をもとに, 検索を行うことも重要な課題となっている。このような文書検索のシステムを構成する場合に基礎となるものが, 文書の属性をもとにして文書を分類する, 文書分類である。

文書分類の手法として, これまで形態素解析などの自然言語処理により単語をキーワードとして抽出し, これらの出現頻度をもとにした特徴ベクトルを求め, この特徴ベクトルに関するクラスタ (特定の特徴をもった文書の集合) 重心の計算と, 分類するべき文書の属性との距離の計算を用いる方法が用いられている。特徴ベクトル空間モデルでは, 文書に含まれる単語 (キーワード) を出現頻度応じた重み付けをして特徴ベクトルを構成し, この類似度によりクラスタ分析を実施する方法を用いる。更に, これを一般化した方法として, 形態素解析に先立つ表現形式である, 単語レベルでの情報抽出を行う n-gram などの方法がある。

このように文書分類を行うための要素分解の手法は, さまざまに提案されているが, しかしながら従来手法では, これを構成的に再現し, 文書分類や検索の方法に用いることについての研究は, 少ないのが現状である。また, 特徴

ベクトルによる分類方法では, キーワードの出現頻度だけが重視され, キーワード間の出現順序が無視されている問題がある。一般に, 情報の出現順序は, 文章の終止状況や語幹による置き換えなどの効果を見る上で重視されている。このようなことから, 単純なベクトル空間の距離ではなく, 一般化された関数として文書間の距離を定義することが望ましい。また, 分類された文書の特徴を特徴ベクトルなどの数値ではなく, 言語的に表現できることが望ましい。

本論文では,GP 手法を用いて文書分類を実現するシステムを提案し, 実際のデータに応用する。GP 手法は, これまで関数近似や, エージェントシステムにおける知識表現, 時系列セグメント認識と時系列予測などへと適用され, 有用性が示されている。本論文では, この手法を文書のクラスタ分類へと拡張している。具体的には, 形態素解析を実施したあとで得られる文書に関するキーワードの出現を考慮して, 解析を行う方法である。特徴ベクトル法では基本的にクラスタの代表ををを1つだけ求めて, この代表値との距離でクラスタを決めているが, 本論文では, 分類のための非線形関数を GP 手法により近似することにより, より柔軟なシステム構成が可能となっている。

GP 手法の基本は, 学習により文書を分類する関数を近似する方法であり, すでにクラスタへの所属が判明している文書を学習データとして用い, この文書を特徴づける変数により記述される関数を多数の GP 手法における個体として与えおき, 安定的に検出されるクラスタが発見されるまで適合度に応じて遺伝的操作を繰り返す方法である。次に, クラスタ所属が未知である文書の特徴を入力として与えた場合に, 関数値が最大となるクラスタに所属する決定を行う。

本論文で示す GP 手法に基づく文書分類システムの利点として GP 手法を適用し, 個体の集合として複数のルールの集合として構成し, 性能向上をはかり, 特徴付けることによりクラスタに含まれるさまざまな変動に対応する分類や推論が可能となることがある。

応用例として, 経済関係の記事のデータなどを用いた分類問題と特徴的な文書の検索題への適用を示す。

9.2 テキスト分類とキーワード抽出

まず, 最初にテキスト分類に基礎的な手法である特徴ベクトルによる方法について述べる。最初に, 形態素解析などの手法により, 文書から語 (words) を切り出す。形態素解析などの手法により, 文書からのキーワード (以下では混乱がない限り, 語と呼んでおく) が抽出される。この語の抽出基準に関しては, 従来からの tf-idf(text frequency-inverse document frequency) が用いられている。文書 k の語 i についての重みは, 次に示す式で計算される。

$$a_{ik} = f_{ik} \log \frac{N}{n_i} \tag{6}$$

ここで N は文書の総数であり,f_{ik} は文書 k に現れる語 i の回数であり,n_i は少なくとも語が 1 回は現れる文書の数である。

この指標に, あるしきい値を設定することにより, 文書分類であまり意味をなさない語を除去することができる。すなわち, 語として用いるものを決定する作業においては, すでに分類が確定している文書を対象として語登録が未定である単語との間の相互情報量である a_{ik} を用いて判断するなどの方法が用いられる。

k 番目の文書の特徴ベクトルは,tf-idf 値を文書の長さ L で正規化したものを要素として構成される。

$$d_k = (a_{1k}, a_{2k}, ..., a_{in}) \tag{7}$$

この場合, 特徴ベクトルを用いて距離を定義する方法により, いくつかのバリエーションが発生することや, 重みを単語の出現頻度に応じて構成する場合に, 頻度の小さい単語を無視するケースを回避する適応的な手法が開発されている。

分類が未知である文書についてカテゴリを定める方法は, 通常のクラスタ分析に類似した方法を用いる。すなわち, カテゴリごとの特徴ベクトルの代表値 (例えばカテゴリに属する学習サンプルの特徴ベクトルの重心) などを求め, この代表値と分類が未知である文書の特徴ベクトルとの差異が最小となるクラスタに分類する方法である。

しかしながら, 特徴ベクトルによる分類方法では, 語の出現頻度だけが重視され, 語間の出現順序が無視されている問題がある。一般に, 情報の出現順序は, 文章の終止状況や, 語幹による置き換えなどの効果を見る上で重視され

ている。このようなことから, 単純なベクトル空間の距離ではなく, 一般化された関数として文書間の距離を定義することが望ましい。また, 分類された文書の特徴を特徴ベクトルなどの数値ではなく, 言語的に表現できることが望ましい。

9.3 GPによるクラスタ分類システムの構成

本論文で示すGPによる文書のクラスタ分類システムにおいては, 基本的にGPよる非線形の分類関数を推定する方法を, 用いている。この方法は, 株価セグメントを認識するシステムなどの応用され, 有効であることが確認されている。

本論文では, 文書分類の特徴ベクトルの要素を変数とする非線形関数を構成し, この関数値によって所属するクラスタを推定する方法を用いる。すなわち, 学習データを用いてクラスタごとに分類関数を構成し, 変数に値を代入した場合に, その値が最大となるクラスタへ所属すると判断する。

いま, 特徴ベクトルの要素について k 番目の語に対応する変数を v_k とする。a_{ik} などは, この変数 v_k の値である。変数 v_k のベクトル $v=(v_1, v_2, ..., v_K)$ に関する関数 $f_c(v)$ をクラスタ $c=1,2,3,...,C$ ごとに仮定する。GP手法においては, 学習データとしてクラスタ c に属する文書の特徴ベクトルをすべての関数に代入したとき, クラスタの関数値 $f_c(v)$ が最大となるように学習を進める。GP手法により関数を最適化する方法の概要を, 以下に示す。

学習データ

システムの概要を図4に示している。この図の中で, 分類に用いる入力データは, 文書に含まれる語に関する特徴ベクトルである。文書分類のシステムにおいては, 分類されるべき排他的なクラスタは複数存在すると仮定し, これらのそれぞれに,1つの木構造の近似関数の集合 (これをGPにおける呼び方にならって, 個体プールと呼ぶ) が対応している。図4上部に示す学習フェーズ (Learning Phase) では, それぞれのクラスタ $c(i=1\sim C)$ ごとにこのような学習データが準備されていることを意味している。すなわち, 外的基準として文書クラスタが分かっていると仮定し, この文書クラスタに相当する特徴ベクトルを学習データとして準備しておく。

個体プールの生成

最初に, 学習データを用いて分類すべき文書クラスタごとに推定する関数の近似形 (GPにおける個体に対応する) を求める。この近似にGP手法を用いる。なお, GPにより生成された近似関数は1つではない。学習データには, ノイズにより変形されたデータや, 基本形からずれたデータも存在するので, 1つの個体で1つのクラスタのすべての特徴ベクトルの特徴を表現することはできない。このような理由から, 相対的に近似度が高い個体を複数選択しておいて, 次の段階の最終的な分類に利用する。多変量解析における判別関数の構成と同様に, 特徴ベクトルの関数を仮定した場合に, それぞれのクラスタに属する特徴ベクトル $v=(v_1, v_2, ..., v_K)$ を与えた場合にだけ関数 $f_c(v)$ が大きくなるように関数 $f_c(v)$ の近似を行っていく。しかしながら, 判別関数を構成する場合と異なり,GP手法においては, 関数を複数準備しておき (これを個体とよぶ), 個体の中で関数の数値が大きなものだけが個体プールに残り, 更に, これを用いて関数近似を改善する方法を用いている。これを示したのが図4上部に示す学習フェーズ (Learning Phase) である。バリエーションのある学習データを, 繰り返し学習に用いるので, 通常の観測データを用いた関数近似の場合と異なり, 適合度の最高値は単調には増加しない。従って, 十分なバリエーションの個体が準備されたと考えられる世代まで, GPを繰り返す方法を用いている。

文書クラスタの推定

以上のような学習を適用して, 個体プールを準備しておく。次に, クラスタ所属が未知である文書の特徴ベクトル $v=(v_1, v_2, ..., v_K)$ を入力した場合に, このクラスタを決定する必要がある。これを示したのが図4下部のカテゴリ化フェーズ (Categorization Phase) である。この場合, すべてのクラスタ k の個体プールの個体 j である関数 $f_c(x)$ に対して特徴ベクトル v を入力変数として代入し, その関数値が最大となる個体がプール c に属している場合, この文書のクラスタを c であると推定する。

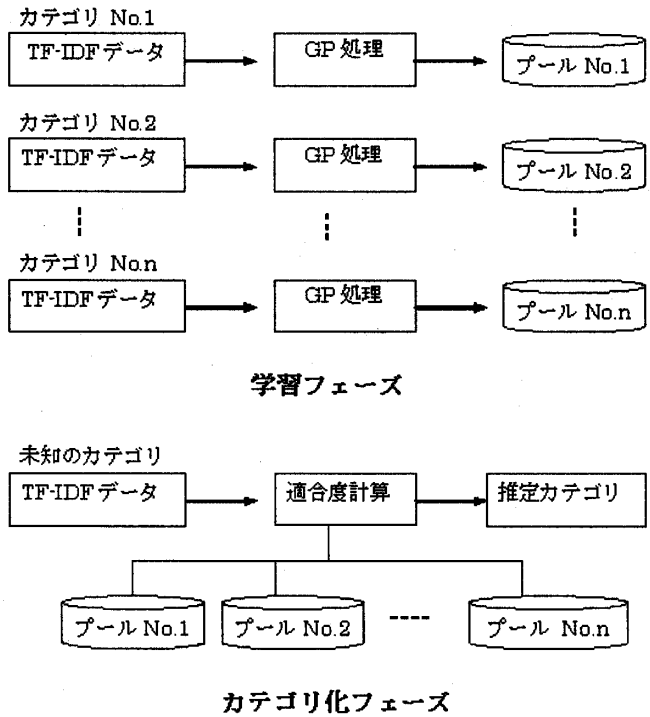

図 4: 文書分類システムの構成概要

9.4 文書分類シミュレーション

以下では, 実際に存在する文書データを用いて, 本論文で示すシステムによる文書分類の性能を評価する。文書は Yahoo!ニュースカテゴリよりクローラソフトを用いて採取・形態素分析を行い、キーワードを得た。これらのニュース文章の一部はあらかじめカテゴリ「科学」「国際」「政治」「経済」「スポーツ」に分けられており、GP により分類ルールを学習・作成する。このルールを用いて、残りのニュース文章を分類する。性能比較のため、顧客情報による購入ブランド分類実験と同様, 余弦法を合わせて行う。余弦法を行う場合, 文章の特徴ベクトルの次元はキーワード数の数と等しくなるため、計算コストが非常に高い。
文書カテゴリ数:5
文章数:350
キーワード数:4556

GP 適用の条件は, 以下のようである。
個体の長さ:20
個体に含まれる演算子:AND,OR
個体プールの中の個体数:1000
GP 適用回数:400
表 8 の結果から分かるように,GP による文章分類の性能は余弦法を大きく上回っており、また文書に登場するキーワードの一部しか計算に使用していないため、計算コストも低い。

表 8. 文章カテゴリ推定の結果

文書カテゴリ	余弦法	GP ルール
科学	39%	75%
政治	70%	88%
経済	69%	71%
国際	55%	75%
スポーツ	57%	70%

10 むすび

本論文では, インターネット・マーケティングの現状と, 課題について述べるとともに,GP 手法による実現ツールの開発について示した。論文の後半では, われわれが開発した GP の手法に基づく顧客管理・文書管理のシステムについて, その基本原理といくつかの実際的に応用例について示した。

今後の課題として, われわれの提案するシステム・ツールを, 実際にインターネット・マーケティングにおいて適用することがあり, 検討を進めていく予定である。

参考文献

[1] W.Hanson,Pronciple of Internet Marketing, South-Western College Publishing,2000.

[2] G.Piatetsky and W.J.Frawley, " Knowledge discovery in database:An overview,"in Knowledge Discovery in Database, AIII/MIT Press, 1991.

[3] A.A.Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer-Verlag, 2002.

[4] B.Baesens, R.Setino,C.Mues and J.Vanthienen, " Using neural netowrk rule extraction and decision tables for credit-risk evaluation", Management Science, vol.49,no.3,pp.312-329,2003.

[5] S.Dutta and S.Shekhar, "Bond rating:A non-conservative application of neural networks," discussion paper, Computer Science Division, University of California, Berkeley ,1989.

[6] 李 鋼浩, 時永祥三, " ニューラルネットワークによる経営情報解析-倒産分析と時系列解析," 経営情報学会論文誌,vol.1,no.2,pp.32-43,1991.

[7] J.R.Quinlan, C4.5 programming for Machine Learning, Morgan Kaufmann, Chambery, France ,1993.

[8] M.W.Craven,J.W.Shavlik, " Extracting tree-structured representations of trained networks, D.Touretzky, M.Mozer,M.Hasselmo,ed. *Advances in Neural Information Processing Systems (NIPS)*,vol.8,pp.24-30,MIT Press, Cambridge,MA.

[9] S.Tokinaga,J.Lu and Y.Ikeda, " Neural network rule extraction by using the Genetic Programming and its applications to explanatory classifications," IEICE Trans.Fuadamentals, vol.E88-A,no.10,pp.2627-2635,2005.

[10] M.L.Wong and K.S.Leung,Data Mining Using Grammer Based Genetic Programmign and Applications,Kluwer Academic Publisher, London,2000.

[11] Y. Ikeda and S.Tokinaga, "Approximation of chaotic dynamics by using smaller number of data based upon the genetic programming," IEICE Trans.Fundamentals, vol.E83-A,no.8, pp.1599-1607, 2000

[12] Y.Ikeda and S.Tokinaga, " Controlling the chaotic dynamics by using approximated system equations obtained by the genetic programming," IEICE Trans.Fundamentals, vol.E84-A, no.9, pp.2118-2127,2001.

[13] 矢加部正幸, 時永祥三, 遺伝的プログラミグを用いたCNNによる拡散モデルの近似と同期化への応用, 電子情報通信学会論文誌, volE85-A, no.5,pp.548-559, 2002.

[14] 池田欽一, "共進化GPによるカオス常微分システムの推定," 電子通信学会論文誌, vol.E85-A,no.4, pp.424-433, 2002.

[15] X.Chen and S.Tokinaga, "Approximation of chaotic dynamics for input pricing at service facilities based on the GP and the control of chaos," IEICE Trans.Fundamentals,vol.E85-A, no.9,pp.2107-2117,2002.

[16] 陳 暁栄, 時永祥三, "G共進化GPを用いたマルチエージェントシステムの構成とその人工市場分析への応用", 電子情報通信学会論文誌, volE86-A,no.10,pp.1038-1048, 2003.

[17] Y.Ikeda and S.Tokinaga, " Chaoticity and fractality analysis of an artificial stock market by the multi-agent systems based on the co-evulutionary Genetic Programming", IEICE Trans.Fundamentals, vol.E87-A,no.9,pp.2387-2394, 2004.

[18] 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.J88-A,no.7,pp.803-813,2005.

[19] 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.J88-A,no.11,pp.1258-1271 ,2005.

[20] 池田欽一, 時永祥三, "GPによる学習を基礎としたマルチエージェント・システムによるプライシング時系列のカオス性分析とその応用", 電子情報通信学会論文誌, 採録決定済み

[21] 池田欽一, 陳暁栄, 時永祥三, "GPによる学習を基礎としたマルチエージェント・システムによるプライシング時系列のカオス性分析とその応用", 情報処理会論文誌, 採録決定済み

[22] J.R.Koza,Genetic Programming,MIT Press, 1992.

[23] J.Koza: " Genetic programming:A paradigm for genetically breeding populations of computer programs to solve problems", Report No.STAN-CS-90-1314, Dept.of.Computer Science Stanford University, 1990

[24] J.Koza, " Evalution and subsumption using genetic programming", Proc of the First European Conference on Artificial Life, MIT Press, 1991.

[25] J.R.Koza,Genetic Programming II:Automatic Discovery of Reusable Programs,MIT Press, 1994.

[26] M.J.Keith and M.C.Martin, " Genetic programming in C++: Implementation issues", in (ed) K.E.Kinnerar,Jr.,Advance in Genetic Programming MIT Press, 1994.

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

[28] Standard & Poor's, "Corporate rating criteria," http://www.coropratecriteria.standardand poors.com,2005 文書分類

[29] G.Salton and M.J.McGill,An Introduction to Modern Information Retrieval,McGraw=Hill,1983.

[30] F.Sebastiani,Machine learning in automated text categorization ," ACM Computing Survey, vol,34 ,no.1,pp.1-47,2002.

[31] H.Hirsch,M.Saeddi and R.Hirsch, " Evolving text clasification with Genetic Programming,"pp.309-317, in (eds) M.Keijzer et.al. Genetic Programming,Springer, 2004.

[32] CD 毎日新聞

高木　昇 (九州産業大学・商学部・商学科・助教授)
時永 祥三 (九州大学大学院経済学研究院・教授)