(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4916614号
(P4916614)
(45)発行日　平成24年4月18日(2012.4.18)　(24)登録日　平成24年2月3日(2012.2.3)

(51)Int.Cl.　F I
G06N 3/00 (2006.01)　G06N 3/00 550C

請求項の数 1　(全 60 頁)

(21)出願番号	特願2000-615965 (P2000-615965)
(86)(22)出願日	平成12年4月19日 (2000.4.19)
(65)公表番号	特表2002-543538 (P2002-543538A)
(43)公表日	平成14年12月17日 (2002.12.17)
(86)国際出願番号	PCT/US2000/010425
(87)国際公開番号	WO2000/067200
(87)国際公開日	平成12年11月9日 (2000.11.9)
審査請求日	平成19年4月18日 (2007.4.18)
(31)優先権主張番号	60/131,804
(32)優先日	平成11年4月30日 (1999.4.30)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	09/466,041
(32)優先日	平成11年12月17日 (1999.12.17)
(33)優先権主張国	米国 (US)

前置審査

(73)特許権者　390023674
イー・アイ・デュポン・ドウ・ヌムール・アンド・カンパニー
E. I. DU PONT DE NEMOURS AND COMPANY
アメリカ合衆国、デラウエア州、ウイルミントン、マーケット・ストリート　１００７
(74)代理人　100127926
弁理士　結田　純次
(74)代理人　100140132
弁理士　竹林　則幸

最終頁に続く

(54)【発明の名称】実験データの分布状階層的発展型モデリングと可視化の方法

(57)【特許請求の範囲】
【請求項１】
増幅ＤＮＡフラグメントの同定に関連するグローバルな情報コンテントを有するフィーチャー集合を選択する、コンピューターにより実行される方法であって、多数の入力データ点と対応する出力データ点が取得されてデータ集合が規定され、該取得された入力及び出力データ点が記憶装置内に記憶され、該フィーチャー集合およびデータ集合はＤＮＡ増幅プロセスの間に生成される方法に於いて、
（ａ）複数のフィーチャー部分空間を創るが、各前記フィーチャー部分空間が該データ集合からのフィーチャー集合を含むように、該創る過程と、
（ｂ）該データ集合の該入力を量子化するが、該入力が値の範囲を有し、それは該値の範囲を部分範囲に分け、それにより前記フィーチャー部分空間を複数のセルに分けることによりするよう、該量子化する過程と、
（ｃ）少なくとも１つのローカルセルのニシのエントロピーＥを計算し、ニシのエントロピーＥの補数としてローカルエントロピーＷ（Ｗ＝１－Ｅ）を規定し、そして各セルのローカルエントロピーＷの加重和を計算することによって、各フィーチャー部分空間のグローバルエントロピーを計算する過程と、
（ｄ）該グローバルエントロピーに基づき、グローバルな情報コンテントを有する少なくとも１つのフィーチャー集合を選択する過程とを具備する、方法。
【発明の詳細な説明】
【０００１】

【発明の分野】
本発明は”対象（objects）”の階層（hierarchy）、例えば、フイーチャー（features）、モデル（models）、フレームワーク（frameworks）、そしてスーパーフレームワーク（super-frameworks）、を創るために、データの画像的表現の概念を情報理論（information theory）からの概念と組み合わせる。本発明はシステムの実験型モデルを、前に取得されたデータ、すなわち、該システムへの入力と該システムからの対応する出力を表すデータ、に基づいて創る方法と機械可読記憶媒体（machine readable storage medium）とに関する。次いで該モデルは次の取得入力からシステム出力を精確に予測するため使われる。本発明の方法と機械可読記憶媒体は情報理論と熱力学の原理に基づく、エントロピー関数を使用し、該方法は複雑な、多元処理（nulti-dimensional process）のモデリングに特に好適である。本発明の方法はカテゴリー的モデリング（categorical modeling）、すなわち、出力変数が離散的状態（discrete states）をとる場合、及び定量的モデリング、すなわち、出力変数が連続的な場合、の両者に使用出来る。本発明の方法は、外見には混乱したシステムであるように見えるものの下にある順序、又は構造を顕わすために、データ集合の最適表現、すなわち最も情報豊富な表現（most information-rich representation）を同定（identifies）する。発展型プログラミング（evolutionary programming）の使用は最適表現を同定する１方法である。該方法は多元的フイーチャー空間（multi-dimensional feature spaces）の情報コンテント（information content）を特徴付ける中でローカル及びグローバルの両情報メザー（both local and global information measure）のその使用により際だっている。実験はローカル情報メザーがモデルの予測能力（predictive capability）を支配することを示した。かくして、全体のデータ集合上でのグローバルな最適化を主として使う、多くの他の方法と対照的に、本方法はグローバルに影響されるが、ローカルに最適化される技術、として説明出来る。

【０００２】
【発明の技術的背景】
情報理論
システムの情報コンテントを説明するためにエントロピー関数（entropy function）を使用する思想は、彼のパイオニヤ的業績、１９４８年発行の、ベルシステムテクニカルジャーナル（Bell System Technical Journal）、２７，３７９－４２３，６２３－６５６、”通信の数学的理論（A Mathematical Theory of Communication）”でシー．イー．シャノン（C. E. Shannon）により初めて導入された。シャノンは統計力学での対応する定義と形式的に同様なエントロピーの定義が起こり得るイベントの総体（ensemble）内での特定のイベントの選択から得られる情報を測定するため使用出来ることを示した。シャノンのエントロピー関数は下記で表され、

【０００３】
【数１】

$$H(p_1, p_2 \cdots p_n) = \sum_{k=1}^{n} p_k \ln p_k$$

【０００４】
ここでp_kは第ｋ番目のイベントの発生確率を示し、ユニークに下記３条件を満足する、
１．Ｈ（p_1，．．．，p_n）はｋ＝１，．．．，ｎでp_k＝１／ｎで最大となる。これは均一な確率分布が最大エントロピーを有することを意味する。加えて、H_{max}（１／ｎ、１／ｎ，．．．，１／ｎ）＝ｌｎ　ｎ。従って、均一確率分布のエントロピーは起こり得る状態の数と共に対数的に縮尺（scales）する。
２．Ｈ（ＡＢ）＝Ｈ（Ａ）＋H_A（Ｂ）ここでＡとＢは２つの有限スキーム（finite schemes）である。Ｈ（ＡＢ）はスキームＡとＢの全エントロピーを表し、H_A（Ｂ）はスキームＢを与えられたスキームＡの条件的エントロピーである。該２つのスキーム分布が相互に独立の時、H_A（Ｂ）＝Ｈ（Ｂ）である。

３．Ｈ（p_1，p_2，．．．，p_n、０）＝Ｈ（p_1，p_2，．．．，p_n）。スキーム内の発生確率ゼロのどんなイベントもエントロピー関数を変化させない。

【０００５】

シャノンの仕事は１次元の電気信号の情報コンテントを説明することに向けられた。１９９８年に、ケンブリッジ大学プレス（Cambridge University Press）で発行された彼の本、フイッシャー情報からの物理学：ユニフイケーション（Physics from Fisher Information: A Unification）で、ロイフリーデン（Roy Frieden）は”シャノンエントロピー（Shannon Entropy）”を全体のデータ集合間のグローバルな情報メザーとして説明している。”フイッシャーエントロピー（Fisher entropy）”として知られる、代わりの情報メザーも又データ集合間のローカルな情報の測定量としてフリーデンにより説明されている。数学的モデル化で、フリーデンはフイッシャーエントロピーが物理的法則を発見するために特に好適であることを最近示した。

【０００６】

より最近に、テー．ニシ（T. Nishi）はどんなデータ集合にも適用出来る、正規化された”情報エントロピー”関数を規定するために該シャノンのエントロピー関数を使用した。１９９１年、京都、３２５、材料の機械的挙動に関する国際会議論文集（Proceedings of the Intenational Conference on 'Mechanical Behaviour of Materials VI'）、ハヤシ、テー．及びニシ、テー．（Hayashi, T. and Nishi, T.）著、”ポリマーアロイの形態学と物理的特性（Morphology and Physical Properties of Polymer Alloys）”、参照。１９９２年発行、高分子論文集（Kobunshi Ronbunshu）、４９（４）、３７３－８２、ハヤシ、テー．、ワタナベ、エイ．、タナカ、エイチ．及びニシ、テー．（Hayashi, T., Watanabe, A., Tanaka, H. and Nishi, T.）著、”３成分不相溶性ポリマーアロイの形態学と物理的特性（Morphology and Physical Properties of Three-Components Incompatible Polymer Alloys）”参照。

【０００７】

ニシの定義は次ぎの様に抄録されるが、ｎのデータ要素（data elements）を有するデータ集合（data set）Ｄ＝｛d_1，．．．，d_n｝を考える。もし全要素の和d_{tot}が次の様に定義されるならば、

【０００８】

【数２】

$$d_{tot} = \sum_{i=1}^{n} d_i ,$$

【０００９】

d_{tot}は、

【００１０】

【数３】

$$f_i = d_i / d_{tot} \quad \forall\ i \in \{1, \ldots, n\}.$$

【００１１】

の様に該データ要素の各々を正規化（normalize）するため使用出来る。

次いで、情報エントロピー関数（informational entropy function）、Ｅを次の様に規定することが出来る、

【００１２】

【数４】

$$E = (\sum_i f_i \ln f_i) / \ln(1/n) .$$

【００１３】
該エントロピー関数Ｅはそれが０と１の間に正規化される有用な特性（property）を有する。f_i＝１／ｎの、完全に均一な分布（perfectly uniform distribution ）は１のＥ値となる。該分布がより不均一になるにつれ、Ｅの値は低下し漸近的にゼロに近付く。該ニシの情報エントロピー関数Ｅの顕著な利点はそれが分布の形状に無関係にどんな分布の均一性も特徴付けることである。対照的に、普通使用される”標準偏差（standard deviation）”はガウス分布（Gaussian disribution）用でのみ標準的統計（standard distribution）に入ると通常解釈される。
【００１４】
ニューラルネットワーク（neural networks）、統計的回帰（statistical regression）、決定木法（decision tree methods）の様な従来技術の方法は或る本質的限定を有する。ニューラルネットワークと他の統計的回帰方法はカテゴリー的モデリングに使用されて来たが、それらは、該ネットワークのノード内で使用される連続非線形シグモイド関数（continuous non-linear sigmoid function）のために、定量的モデル化に遙かにより適合し、より良く動作する。決定木は、連続的出力値に関する精確な定量的予測をする能力に欠けるためにカテゴリー的モデリングに最も良く適合している。
【００１５】
【発明の概要】
本発明は情報エントロピーの概念を一般化し、それらの概念を多次元データ集合へ延長している。特に、シャノンにより表明された情報エントロピーの定量化は修正され、１つ以上の入力、又はフイーチャー、と１つ以上の出力とを有するシステムから得られたデータに適用される。情報豊富（information-rich）でありかくして該システム出力（含む複数）の予測に有用なデータ入力の種々の部分集合（subset）、又はフイーチャーの部分集合を同定（identify）するためにエントロピー定量化（entropy quantification）が行われる。又該エントロピー定量化は情報豊富な種々のフイーチャー部分集合内で領域（region）、又はセル（cell）を同定する。該セルは固定的又は適合的なビニング過程（binning process）を使用してフイーチャー部分空間内で規定される。
【００１６】
入力組み合わせ（input combination）、又は特徴組み合わせ（feature combination）、はフイーチャー部分空間を規定する。該フイーチャー部分空間は２進ビット記号列（binary bit string）により表され、ここでは遺伝子（genes）として引用される。遺伝子はどの入力が特定部分空間にあるかを示し、従って特定の部分空間の次元数（dimensionality）は該遺伝子数列（genes sequence）の”１”のビットの数により決定される。望ましい情報特性を有する部分空間に対応するそれら遺伝子を同定するために全てのフイーチャー部分空間の情報豊富さがエグゾースチブ（exhaustively）に探索される。
【００１７】
起こり得る部分空間（possible subspace ）の全数が少なければ、エグゾースチブな探索が最も情報豊富な部分空間を同定する好ましい方法であることは注意すべきである。多くの場合、しかしながら、起こり得る部分空間の数は全ての起こり得る部分空間をエグゾースチブに探索することが計算的に非現実的である程充分大きい。それらの状況では、該部分空間は遺伝子数列を操作する遺伝的アルゴリズムを使用して探索されるのが好ましい。すなわち、遺伝子は望ましい情報特性を有するフイーチャー部分空間の集合を進化させるよう組み合わされ及び／又は選択的に突然変異（mutated）させられる。特に、該遺伝的フイーチャー部分空間進化過程（evolution process）用の適応度関数（fitness function）はその特定の遺伝子により表されるフイーチャー部分空間用情報エントロピーのメザー（measure）である。情報コンテントの他のメザーは該出力に関する該部分空間の均一度を示す（measure）。これらのメザーは分散（variance）、標準偏差、又は或るしきい値を越える指定出力依存確率を有するセルの数（又はセルのパーセンテージ）の様な発見的方法（heuristics）を含む。これらの情報的メザーは望ましい情報特性、すなわち高い情報コンテントを有する遺伝子、又は部分空間を同定するために使用されてもよい。加え

て、決定木ベースの方法が使用されてもよい。これらの代替えの方法はエグゾースチブな探索を行う時望ましい部分空間を同定するため使用されてもよい。

【００１８】

好ましい実施例では、ここではグローバルエントロピーと呼ぶ、該フイーチャー部分空間エントロピーは、該部分空間内のセルのエントロピーメザーの加重平均を計算することにより決定されるのが好ましい。出力特定的エントロピーメザーも又使用されてもよい。セルエントロピーはここではローカルエントロピーと呼ばれ、修正されたニシのエントロピー計算を使用して計算される。

【００１９】

実験型モデルが次いで階層的な仕方で創られるが、それは、高い情報コンテントを有するよう決定されたフイーチャー部分空間の組み合わせを調べることによる。フイーチャー部分空間は、テストデータ（既知の対応出力を有するサンプル入力データ点）を使用する高精度の予測を提供するフイーチャー部分空間の組み合わせを見出すためにエグゾースチブな探索技術を使用して選択されそしてモデル内へ組み合わされる。該モデルは又遺伝的アルゴリズムを使用して発展させられてもよい。この場合、該モデル遺伝子はどのフイーチャー部分空間が使用されるかを指定し、該モデル遺伝子の長さは望ましい情報特性を有するとして前に同定されたフイーチャー部分空間の数により決定される。該モデル発展過程で使用される該適応度関数は考慮下の特定モデルの予測精度であるのが好ましい。

【００２０】

本発明の１側面に依れば、次ぎに取得される入力からシステム出力を精密に予測するため、該システムへの対応する入出力を表す、前に取得されたデータに基づきシステムの実験型モデルを創る方法が提供される。該方法は、

（ａ）該システムへの多数の入力と対応する該システムからの出力とからデータ集合を取得する過程と、

（ｂ）該前に取得したデータ集合を、少なくとも１つのトレーニングデータ（training data）集合と、少なくとも１つのテストデータ（test data）集合と、そして少なくとも１つの検証データ（verification data）集合とにグループ分けする過程を具備しており、該集合は相互に一致してもよく、或いは前に取得したデータの排他的（exclusive）又は非排他的（non-exclusive）部分集合であってもよく、該方法は又、

（ｃ）高いグローバルエントロピー加重（weights）を有する複数のフイーチャー部分空間を、

（ｉ）前記トレーニングデータ集合からフイーチャー部分空間を規定する複数の入力を選択する過程と、

（ｉｉ）固定的か又は適合的か何れかの量子化方法（quantization）により、各入力範囲を部分範囲（subrange）に分けることにより該フイーチャー部分空間をセルに分ける過程と、

（ｉｉｉ）ローカルセルラーエントロピー加重による加重平均か、又は出力特定的エントロピー加重による加重平均か何れかを形成することにより、グローバルエントロピー加重を決定する過程と、

により決定する過程と、

（ｄ）オプション的に、高いエントロピー加重を有する該決定されたフイーチャー部分空間内での各入力発生の頻度を調べ、削減された次元数データ集合を規定するために最も頻繁に発生するそれらの入力のみを保持する過程と、そしてその後過程（ｃ）を繰り返す過程と、

（ｅ）オプション的に、該削減された次元数フイーチャーデータ集合を規定するようにシステム入力から最も精密にシステム出力を予測する最適又は最適に近い次元数と最適又は最適に近い量子化条件を決定するために、複数の量子化条件下で該削減された次元数データ集合の複数の該次元（例えば、該次元の幾つか、又は全て）上でエグゾースチブに探索する過程と、

（ｆ）前記データ集合上のシステム入力からシステム出力を最も精密に予測する高いグロ

ーバルエントロピー加重（例えば、フイーチャーデータ集合の部分か、又は全体か何れか）を有する該決定されたフイーチャー部分集合の組み合わせを決定する過程と、
（ｇ）テストデータ集合上でシステム入力からシステム出力を最も精密に予測する削減された次元数のフイーチャーデータ集合に部分集合（例えば、削減された次元数のフイーチャーデータ集合の部分か、又は全体かの何れか）を決定する過程とを具備している。
【００２１】
大きなデータ集合用には、該モデル創生過程（ｂ）－（ｇ）は、次いで最適モデルのグループを見出すために種々のトレーニング及びテストデータ集合上で繰り返されてもよい。この最適モデルのグループはそれらのモデルから生じる１つ以上の予測を開発するために新しいデータについて”ポール（polled）”されてもよい。これらの予測は、例えば、勝者１人占め（winner-takes-all）の投票ルールに基づいてもよい。システム入力から最も精密にシステム出力を予測する最適モデルのグループの部分集合は次いで次の様に決定される。テストデータ集合の入力がモデルの選択された部分集合のグループの各モデルに従属させられ（ランダムに選択されてよい）、各部分集合で予測された出力は各テストデータ出力と比較される。該部分集合で予測された出力の計算過程は（ｂ）－（ｅ）｛又はオプションとして（ｂ）－（ｇ）｝と同様な仕方で行われ、そこでは個別のモデル出力予測値を入力として、実際の出力値を出力として使用して新しいトレーニング及びテストデータ集合が創られる。この過程はモデルの多数の選択された部分集合グループ用に繰り返されてもよい。モデルの該選択された部分集合グループは次いで、”フレームワーク”を規定するためにシステム入力からシステム出力を最も精密に予測するモデルの最適部分集合グループを見出すために発展（evolved）させられる。
【００２２】
フレームワーク創生過程は、最適フレームワークのグループを見出すために、モデル創生過程と同様な仕方で更に繰り返されてもよい。最適フレームワークのこのグループは、それらのフレームワークから生じる１つ以上の予測を開発するために新データ上で”ポール”され得る。これらの予測は、例えば、勝者１人占めの投票ルールに基づくことが出来る。システム入力からシステム出力を最も精密に予測する最適フレームワークのグループの部分集合は次いで次の様に決定される。テストデータ集合の入力はフレームワークの該選択された部分集合グループの各フレームワークに印加され、各フレームワーク部分集合で予測された出力が各テストデータ出力と比較される。該部分集合で予測される出力の計算過程は（ｂ）－（ｇ）と同様な仕方で行われ、そこでは個別モデルフレームワークで予測された値を入力としてそして実際の出力を出力として使用して新トレーニング及びテストデータ集合が創られる。この過程はフレームワークの多数の選択された部分集合グループ用に繰り返される。フレームワークの該選択された部分集合グループはシステム入力からシステム出力を最も精密に予測する、”スーパーフレームワーク”と呼ばれる、フレームワークの最適部分集合グループを見出すために発展させられる。
【００２３】
最適モデル決定過程、最適フレームワーク決定過程、又は最適スーパーフレームワーク決定過程は予め決められた停止条件が達成されるまで繰り返される。該停止条件は、例えば、１）発展型対象の族（family of evolutionary objects）のポーリングから予め決められた予測精度の達成、又は２）予測精度でのインクレメンタルな改善が予め決められたしきい値より低下した時、又は３）予測精度での更に進んだ改善が達成されない時、として規定されてもよい。
【００２４】
分布状階層的発展（Distributed hierarchical evolution）は、モデル、フレームワーク、スーパーフレームワーク他の様な逐次的により複雑に相互作用する発展型”対象”のグループが、逐次的により大量の複雑なデータをモデル化し理解するために、創られる発展型の過程である。
【００２５】
【本発明の詳細な説明】

図１は本発明の方法１００の全体的流れを図解するブロック線図である。この図から評価される様に、実験データから複雑なシステムのモデルを創生するために発展型過程（evolutionary process）が使用される。好ましい方法は、”発展型対象（evolutionary objects）”、例えば、フイーチャー１３０、モデル１４０、フレームワーク１５０、そしてスーパーフレームワーク１６０他、の伸展する階層（extensible hierarchy）を創るために、データ１１０の多次元的表現を情報理論１２０と組み合わせる。該過程は１７０で示した階層的な仕方で更に組み合わせを発生するため続けられ得る。

【００２６】

最初に、フイーチャー部分空間（feature subspace）とも呼ばれる、入力の組み合わせは、初期のランダムに選択されたフイーチャー部分空間プールからエグゾースチブな探索（exhautive search）又は発展型の過程により、同定（identified）される。次いでモデルを創るためにフイーチャー部分空間の最適組み合わせ（optimum combination）が探索されるか又は発展（evolved）させられ、フレームワークを創るためにモデルの最適組み合わせが更に探索されるか又は発展させられ、そしてスーパーフレームワーク他を創るためにフレームワークの最適組み合わせが更に探索されるか又は発展させられる。上記説明のより複雑な発展型対象の逐次的発展は、予め決められた停止条件、例えば、予め決められたモデル性能、が達成されるまで続く。ルールとして、該データ集合（data set）が大きい程、これらの対象のより多くが創られるので、実験型モデル（empirical model）の複雑さは、該入力の、該データが取得された該システムの出力との相互作用の複雑さを反映する。

【００２７】

ここに説明した方法の展開で、幾つかの設計基準（design criteria）が考えられた。該方法が、任意の非線形構造を有するデータ空間（data space）を成功裡に処理することが必要である。該方法が、入力を知って出力を予測する”前向き（foreward）”問題と、出力を知って入力を予測する”逆向き（inverse）”問題との間を区別せず、それによりデータのモデル化と制御の問題を同じ足場（footing）上に置くことも又望ましい。これは該データ集合それ自身の上に最小の追加的モデルジオメトリー（additional model geometry）だけが重ね合わされることを意味する。用語”ジオメトリー（geometry）”は、回帰技術（regression technique）で導入される様な、線形及び非線形の両多様性を含む。対称性（symmetry）もここでは目下のモデリングタスク用に最も情報豊富な（information-rich）入力又は入力の組み合わせを同定する利点を有する。この知識は意志決定及び計画用の最適戦略を開発するため使用され得る。最後に、該方法は、それが事実便利に実施されるために計算的に扱い易い（tractable）必要がある。これらの設計目標を充たすために、幾つかの現在の線形及び非線形な方法が注意深く解析され、共通のテーマが基本的な限定と機会とを同定する目標を用いて要約された。

【００２８】

下記の議論は情報理論及び発展からの概念を使用して１つのモデルの発展の基本的方法を説明することから始まる。より大きい。より複雑なデータ集合を説明するために逐次的により複雑な対象の逐次的で階層的な発展に向かうために該方法を更に伸展させることが次ぎに説明される。データ出力がなくても入力フイーチャークラスター（input feature cluster）を発見する方法の下にある原理の応用が次いで論じられ、それに多次元データ空間内で”情報可視化（information visualization）”を行う方法の説明が続く。ハイブリッドのモデリングスキームを創るために本発明の方法をニューラルネットワーク（neural networks）の様な他のモデリングパラダイム（modeling paradigms）と組み合わせることが次いで詳述される。該説明は、遺伝的プログラミング（genetic programming）の分野と結合された本発明の方法のデータモデル化の取り組みを使用して物理的法則を発見する、新しい取り組みを結論としている。

【００２９】

関心の点として、情報理論からの基本的アイデアは全てのこれらの問題を解くに必要なコアツール（core tools）を提供し、簡単で統合的核（simple, unifying kernel）を該方

法に提供することは述べるに値する。エントロピー（entropy）の概念はデータ空間内の秩序（order）｛又は混乱（disorder）｝の定量的メザー（quantitative measure）を提供する。このメザーは、初期に混乱したシステムからの秩序の発生をドライブする発展型エンジン用の適応度関数（fitness function）として使用され得る。この意味で、情報理論はドライバーを提供し、発展型プログラミングは発見過程をシステム化するエンジンを提供する。最後に、本発明の方法で説明されるパラダイムはデータドライブされている（is data driven）が、それはデータ自身の中の情報コンテント（information content）が予測（prediction）に使用されるからである。かくして、該方法は、下にある数学のその固有の制限を有する数学的モデル化の分野と反対に、実験型モデル化の分野に真正面（squarely）から属する。

データモデリング（DATA MODELING）

情報エントロピーの概念に基づくフレームワークは、入力の集合を与えられたとして１つか又は多数か何れかの出力が予測される必要がある様な、データモデリングの問題に適用されて来た。基本的方法は次の過程から成るが、すなわち

１．データ表現（data representation）又はデータ事前処理（data preprocessing）、

２．セル境界（cell boundary）を規定する固定的又は適合的（adaptive）な方法を使用するデータ量子化（data quantization）、

３．遺伝的発展及び情報エントロピーを使用するフイーチャー組み合わせ選択、

４．システム入力からシステム出力を最も精密に予測するフイーチャーデータ集合の部分集合（subset）の決定である。

１．データ表現

典型的な実験的に得られたデータ集合で、幾つかの”測定”入力と出力とが提供される。各システム入力とシステム出力は、ここでデータ点（data points）と呼ぶ、データ値の入力及び出力のシーケンスを得るようにサンプリングされるか他の仕方で測定される。目標（goal）は該データ点出力を最も精確に予測するために該データ点入力から最大の情報を抽出することである。多くの実システム（real syatem）では、該データ点、又は実際の測定された入力は、それらが該データの適切な表現として留まるに充分な程”情報豊富（information-rich）”である。他の場合は、これはそうでないかも知れず、該データを表現するより適切な”固有ベクトル（eigenvectors）”を創るために該データを変換することが必要かも知れない。共通に使用される変換には特異値分解法（singular value decomposition）｛エスブイデー（SVD）｝、主成分分析法（principal component analysis）｛ピーシーエイ（PCA）｝、部分的最小２乗法（partial least square method）｛ピーエルエス（PLS ）法｝が含まれる。

【００３０】

最も大きい対応する”固有値（eigenvalues）”を有する主成分”固有ベクトル”（eigenvectors）が該データモデリング過程用入力として通常使われる。該主成分選択法には２つの顕著な限定がある。

【００３１】

ａ．該主成分法は入力の分散のみを取り扱い、出力に関する情報は何もエンコードしない。多くのモデリング問題で、モデル化されつつある出力特性に関する最も多くの情報を含む比較的低い固有値を有するのは固有ベクトルである。

【００３２】

ｂ．該ピーシーエイ法は入力の線形変換を行う。これは全ての問題用には、特に入力－出力関係が非常に非線形であるそれら用には最適変換ではないかも知れない。

【００３３】

ここで説明する方法の好ましい実施例では、その組み合わせが”入力フイーチャー（input features）”としても知られる、入力は初期には変換されない。もし次の入力データ集合が、モデル化される必要のある出力に関する充分な情報を現さないならば、上記で説明されたそれらの様なデータ変換が行われてもよい。この戦略を使う主な理由は、変換の形式内に追加的ジオメトリーを課すよりも、可能な所ではどこでも実際のデータを使用する

ことである。この追加的ジオメトリーが取る形式は未知であるかも知れない。加えて、データ変換過程を避けることは該変換過程の計算的オーバーヘッドを避け、かくして、特に非常に大きなデータ集合用の計算効率を改善する。

【００３４】

実際のデータが好ましくは変換なしで使用されるのがよいとは云っても、他の入力よりも情報豊富な入力、又はフイーチャーを同定し、選択することにより次元数（dimensionality）はなお減じられてもよい。これは、入力数が非常に多い時は特に望ましく、最終モデルに起こり得るフイーチャーを全て使用することは非実用的である。データ集合の”次元（dimension）”は入力の全部の数として規定されてもよい。実験型モデルを開発する前に、好ましくは、当面のモデリングタスク用に最も情報豊富なフイーチャーを同定されるのがよい。入力数を減じる、又は該問題の次元数を減じる１つの技術は、少しの情報コンテントしか持たない入力を除くことである。これは入力と、対応する出力と、の相関（correlation）を調べることに依りなされてもよい。しかしながら、好ましくは、次元数削減は、下記で論じる様に、情報豊富と決定されたフイーチャー組み合わせで各入力の発生頻度（each input's frequency of occurrence）を調べることにより行われるのがよい。それで、より少ない発生頻度の入力（less-frequently-occurring inputs）はモデル発生過程から排除されてもよい。

【００３５】

時間変化する又は動的なシステム用では、追加的複雑さが、与えられた何れかの時の出力が、より早期の時の入力と出力との双方にも左右される事実から生ずる。この様なシステムでは、該データ集合の正しい表現が非常に重要である。もし特定時刻の測定出力に対応する入力がその時だけ測定されるならば、該時間遅れ（time lags）（すなわち、入力発生と該結果としての出力発生の間の時間間隔）内に含まれる情報は失われる。この問題を緩和するために、入力の拡張された集合から成るデータ表（data table）が作られるが、そこでは該入力の拡張された集合は入力の現在の集合のみならず多数の前の時刻（at multiple prior times）の入、出力からも成っている。この新データ表は次いで選択された時刻範囲に亘り（spanning a selected time horizon）情報豊富な入力組み合わせ用に解析され得る。

【００３６】

拡張データ表の創生での重要な事項は時間的に如何に遠くまで逆戻って知るかである。多くの場合、これは先験的には知られず、余りに長く早期までの時間間隔｛時間範囲（time span）｝を含めることにより、該データ表の次元数は非常に大きくなる。この事項を処理するために、多数のより短い時間範囲のデータ表が元のデータ表から作られるが、各データ表は過去での与えられた時間間隔から成る。これらのより新しいデータ表の各々の及ぶ時間間隔は重なったり、隣接したり又は分離していてもよい。これらのより小さいデータ表の各々からの最も情報豊富な入力が次いで集められ、該小さなデータ表からの選択された入、出力を含むハイブリッドデータ表を作るよう組み合わされる。この最後のハイブリッド表は、該時間間隔間の起こり得る相互作用が今や含まれるので、次いでデータモデル化過程への入力として使用出来る。

【００３７】

例えば、もし住宅販売レート（home sales rate）が商品製材価格（commodity lumber prices）に影響するが、約２ヶ月の推定時間遅れがあるのでないか、を調査したいならば、この時間遅れを発見するために本発明用には該データ表は入力が出力に２ヶ月先行する対応（matched）した入、出力を要する。これは、実際の時間遅れがどれだけかを発見するために種々の入力が１つの出力に対し異なる遅れを有する１つ以上のデータ表（すなわち、列は入、出力、行は連続した時間）を形成することにより行われ得る。特に、１つの出力はＸ日の製材価格であってもよい。入力がＸ日、Ｘ－１日、Ｘ－２日．．．からＸ－１２０日までの住宅販売レートであるのみならず、Ｘ－１、Ｘ－２．．．からＸ－１２０までからの出力でもある。高い情報コンテントを持つ最も早期の入力が失われないことを保証するために、入力と対応する出力との間の推定時間遅れ（suspected time lag）より

長い時間間隔が選択される。次いで次の表の行はＹ日（例えば、Ｘ＋１又は幾らかもっと後れた日）の製材価格に等しい出力を有し、入力はＹ、Ｙ－１、Ｙ－２，．．．Ｙ－１２０の住宅販売レートであるのみならずＹ－１、Ｙ－２．．．からＹ－１２０日までからの出力でもある。次いで該システムは該出力に影響する入力の組み合わせを同定することにより適当な時間遅れを同定する。２．データ量子化とフイーチャー部分空間内のセル境界一旦適当なデータ表現が確立されると、サンプル点を特徴付けるため使用される各入力で”量子化（quantization）”過程が行われる。入力値の範囲を部分範囲に分ける、すなわち、当該技術で”ビニング（binning）”として公知の、ビン（bins）に分けるために２つの量子化方法が使われるが。該ビニングは与えられたフイーチャー部分空間の各入力で行われるが、そこでは各入力は該部分空間の次元に対応し、それはセルの領域に分けられる与えられたフイーチャー部分空間となる。

【００３８】

最も簡単な量子化法は固定サイズの部分範囲、すなわちビン幅（時には、”固定ビニング（fixed binning）”として知られる）に基づくが、そこでは各入力に付随する値の全体範囲が等間隔又は等サイズの部分範囲又はビンに分けられる。

【００３９】

もう１つの量子化、それは”統計的量子化（statistical quantization）”と呼ばれてもよく、図２Ａで最も良く見られ、ここでは”適合的量子化（adaptive quantization）”と呼ぶが、は値の該範囲を不等サイズの部分範囲に分けることに基づく。もしデータがデータビン２１０により示す様に均一に分布されていれば、該ビンサイズは大体等しい。しかしながら、該データ分布がクラスター（clistered）されるならば、該ビンサイズは、ビン２２０により示される様に、各ビンがデータ点の殆ど等しい数を含むように適合的に調整される。図２Ｂに見られる様に、各部分範囲、又はビンのサイズは、入力範囲を等しい百分位数（percentile）の部分範囲に分け、それらの百分位数を該ビン２４０を作るフイーチャー値の範囲上に射影（projecting）することにより、各入力の累積確率分布（cumulative probability distribution）２３０（又はヒストグラム）に関係付けられてもよい。

【００４０】

この方法で、各入力上のグローバル情報がその入力上で該データを適合的に量子化するため使われる。この方法では、各入力は別々に量子化され、すなわち、量子化は入力毎ベースで行われる。該部分範囲又はビンのサイズ（幅）は与えられた入力内で一般に不均一で、その入力の累積確率分布の形を反映していることを注意すべきである。該部分範囲のサイズは入力から入力へと変わってもよい。適合的量子化（適合的ビンニング）は情報を含まない空の入力の部分範囲を有する確率を減らすが、それはさもないと最終モデル内の情報ギャップとなる。

【００４１】

与えられた入力に対する該部分範囲、又はビンのサイズは部分空間から部分空間へと変わってもよい。すなわち、或る入力は、それらが高い次元の部分空間で現れる時より低い次元の部分空間で現れる時の方がより精細な解像度のビニングを有してもよい。これは或る全体のセルの解像度（セル当たりの点の数）は、データの意味のある量がセル内で一緒にグループ化又はビン化（binned）されるように、望まれる事実のためである。セル数は次元数に指数関数的に比例するので、より高い次元のフイーチャー部分空間は、セル当たりの望ましい平均の点の数を保持するように、個別入力用により粗いビニングを使用する。データ量子化がモデル化の方法のローバストさ用に顕著な意味を有するのは該データの残りからの外れ値の点の偏差の大きさが該量子化（ビニング）過程中に抑制されるからである。例えば、もし入力値が最高部分範囲（ビン）内の上限を越えるなら、それはその値に無関係にその部分範囲（ビン）内に量子化（ビン化）される。

【００４２】

ここで使用される”フイーチャー部分空間”は１つ以上の入力の組み合わせと規定される。フイーチャー部分空間の画像的表現が創られてもよく、それも又簡単に”部分空間”と

してここでは呼ばれる。該部分空間は好ましくは複数の”セル”に分けられるのがよく、該セルは該フイーチャー部分空間を含む入力の部分範囲の組み合わせにより規定される。好ましい実施例では、データ量子化は更に、（前の説明の固定的か又は適合的か何れかの方法を使用して）入力当たりの部分範囲（ビン）の数を規定するか、又は、代わりに、該フイーチャー内のセル当たりデータ点の平均数を規定するか、何れかで指定される。これは適合的量子化法の多次元的拡張と見られる。

【００４３】

図３Ａ、３Ｂそして３Ｃを参照すると、固定サイズのビニングがそれぞれ１，２そして３次元フイーチャー部分空間で示される。該データ集合は各々が４つの入力、又はフイーチャーを有する４つのデータ点、ＤＰ１－ＤＰ４から成る。該データ集合は全ての３つの図で同じである。該データ点はどのフイーチャー（又はフイーチャー組み合わせ）が選択されるかにより特定のセルに分類される。図３Ａでは、もし該１次元部分空間が第３の入力（左端のビットに対応する第１入力を用いて００１０と呼ばれる）を表せば、ＤＰ１とＤＰ４はセルＣ１に分類され（ＤＰ１＝．５、ＤＰ４＝．３）、ＤＰ２とＤＰ３はセルＣ２に分類される（ＤＰ２＝１．２、ＤＰ３＝１．７）。もし、しかしながら、該１次元部分空間が第２入力（０１００）であると取られるなら、ＤＰ２とＤＰ４はＣ１に分類され（ＤＰ２＝．７、ＤＰ４＝．４）、そしてＤＰ１とＤＰ３はＣ２に分類される（ＤＰ１＝１．５、ＤＰ３＝１．９）。

【００４４】

図３Ｂでは、もし該部分空間が第１と第２入力（１１００）により指定されれば、ＤＰ１はセルＣ２に分類される｛ＤＰ１＝（．５、１．５）｝が、なお該第１と第３入力（１０１０）により発生される部分空間ではセルＣ１に分類される。図３Ｃでは、ＤＰ１は第１、第３そして第４入力（１０１１）で規定される部分空間ではセルＣ１に分類され、第１、第２そして第４入力（１１０１）で規定される部分空間ではセルＣ２に分類される。

【００４５】

該入力に基づく該システムの出力の予測で或る精度を有するフイーチャー組み合わせを同定することが望ましい。特定の入力組み合わせ、又はフイーチャー組み合わせは多くのユニークな部分空間を規定することが上記例から分かる。有限数の入力シーケンスを仮定すれば、の部分空間の数は勿論有限であるが、該数は入力数と共に極めて急速に成長する。

【００４６】

フイーチャー選択のタスクは入力－入力の相互作用の可能性により複雑化する。この様な相互作用が存在すれば、個別には情報貧弱な入力が高い情報エントロピーを有する入力の組み合わせを作る相補的な仕方で組み合わされ得る。かくして、入力－入力相互作用の可能性を無視するどんなフイーチャー選択方法もモデル化過程から有用な入力を排除する可能性があり得る。この制限を避けるために、好ましい方法は、入力－入力関係を本質的に含み、該データ内にあるかも知れぬ何等かの非線形性を非常に自然に処理する、情報理論ベースのフイーチャー部分空間を選択する取り組みを使用する。

【００４７】

加えて、該方法は利用可能な部分空間のエグゾースチブ（exhaustive）な探索を含むが、それが好ましくは情報エントロピーのメザーを適応度関数として使う遺伝的発展型アルゴリズム（genetic evolutionary algorithm）を含むのがよい。

３．遺伝的発展と情報エントロピーを使用するフイーチャー部分空間選択

ここで説明する方法は好ましくは”遺伝的アルゴリズム”として公知の比較的最近のアルゴリズム的取り組みを使用するのがよい。ジョンエイチ．ホランド（John H. Holland）｛１９７５年発行、アナーバー、ミシガン大学プレス（Ann Arbor:the University of Michigan Press）、”天然及び人工的システムでの適合（Adaptation in Natural and Artificial Systems）”で｝により定式化され、又デー．イー．ゴルドバーグ（D. E. Goldberg）｛１９８９年発行、アデイソン－ウエズレーパブリッシングカンパニー（Addison-Wesley Publishing Company）、”探索、最適化及び機械学習に於ける遺伝的アルゴリズム（Genetic Algorithms in Search, Optimization and Machine Learning）”で｝及びエ

ム．ミッチェル（M. Mitchell）｛１９９７年発行、エムアイテープレス（M.I.T. Press）、”遺伝的アルゴリズム入門（An Introduction to Genetic Algorithms）”で｝により説明された様に、該取り組みは最適化問題を解く強力で、一般的な方法である。遺伝的アルゴリズムの取り組みは次の様である。

【００４８】

（ａ）問題の解空間（solution space）をＮビット記号列（N-bit strings）の母集団（population）としてエンコードする。ポピュラーなエンコード用フレームワークは２進記号列（binary strings）に基づく。該ビット記号列の集まりは”遺伝子プール（gene pool）”と呼ばれ、個別ビット記号列は”遺伝子（gene）”と呼ばれる。

【００４９】

（ｂ）目前の問題に対する何等かのビット記号列の適応度（fitness）を測定する適応度関数（fitness function）を規定する。換言すれば、該適応度関数は何等かの起こり得る解の良さ（goodness）（又は精度）を測定する。

【００５０】

（ｃ）ビット記号列のランダムな遺伝子プールで最初にスタートする。それを通してより”適した（fit）”ビット記号列が”より適した（fitter）”子供（offspring）の新しいプールを作るために優先的にメートする、選択的再組み合わせ（selective recombination）及び突然変異（mutation）の様な、遺伝子から得られたアイデアを使用することにより、より適したビット記号列の次の世代が発展出来る。”適応度（Fitness）”は情報エントロピーのメザーにより決定される。突然変異の役割は起こり得る解の探索空間を拡張することであり、該解は改善された度合のローバストさ（robustness）を創る。

【００５１】

（ｄ）上記進め方に従う数世代の発展の後、より適したビット記号列のプールとなる。最適解はこのプール内の”最適（fittest）”ビット記号列として選択される。

【００５２】

これらの側面の各々を下記で更に詳細に論じる。

ａ．Ｎビット記号列の母集団としての解のエンコーデイング（Encoding solution as a population of N-bit strings）

最適問題を解くために遺伝適アルゴリズムを使う最初の過程は、ビット記号列として表される解となる方法で該問題を表すことである。簡単な例は４入力と１出力を有するデータベースである。入力の種々の組み合わせが４ビット２進記号列により表される。該ビット記号列１１１１は、全ての入力が該組み合わせ内に含まれる入力組み合わせ、又はフイーチャー部分空間を表す。最左ビットを入力Ａ、第２の最左ビットをＢ、第３の最左ビットを入力Ｃそして最右ビットを入力Ｄと呼ぶ。もしビットが値１に換わるなら、それは対応フイーチャーが該組み合わせ内に含まれるべきことを意味する。逆に、もしビットが値０に換わるなら、それは対応フイーチャーが該組み合わせ内で排除されるべきことを意味する。

【００５３】

同様に、該ビット記号列１０００は唯フイーチャーＡが含まれ、全ての他の入力が排除される入力組み合わせを表す。この方法で、１６の全可能性からのあらゆる起こり得る入力組み合わせは４ビット２進記号列により表される。一般に、もしモデル化されるデータベースにＮ入力があるなら、全ての起こり得る入力組み合わせはＮビット２進記号列を使用して表される。４次元のフイーチャー部分空間を表すサンプルの２進ビット記号列は図４に示される。図４の該ビット記号列はＤビットを有し、その４つだけが”１”のビットである。該”１”のビットは４つのフイーチャーＦ１，Ｆ４，Ｆｉ、そしてＦＤと対応する。該変数ｉとＤは一般化された場合を表すために使用される。更に進んだ例が図３Ａで示されるが、そこでは４入力システムを表し、１つの”１”ビットを有する、４ビット記号列が１次元フイーチャー部分空間に対しコード化する。２つの”１”ビットが図３Ｂに見られる２次元部分空間に対しコード化し、３つの”１”ビットが図３Ｃで見られる３次元部分空間に対しコード化する。

ｂ．ビット記号列の適応度を測定するための適応度関数の規定
最適化問題への解として最適ビット記号列を発展させるために、発展過程をドライブするため使用される定量評価（metric）を規定することが必要である。この定量評価は遺伝的アルゴリズムでは適応度関数と呼ばれる。それは与えられたビット記号列が如何に良く目前の問題を解くかのメザー（measure）である。適当な適応度関数を規定することは該ビット記号列がより良い解へ発展することを保証する重要過程（critical step）である。
【００５４】
上記例では、各４ビット２進記号列は入力の起こり得る組み合わせをエンコードする。入力フイーチャー部分空間は、対応するビット記号列内でオンに換わる入力フイーチャーを使用することにより作られ得る。データベース内のデータはこのフイーチャー部分空間内へ射影され得る。該適応度関数は、該入力フイーチャー部分空間上で出力状態の分布を調べることにより情報豊富さのメザーを提供する。もし該出力状態がこの部分空間上で非常にクラスターされてそして分離されていれば、該対応する入力フイーチャー組み合わせは異なる出力状態を分離することでよい仕事をしているので該適応度関数は高い値となる。逆に、もし全ての出力状態が該部分空間上にランダムに分布されているならば、該対応する入力フイーチャー組み合わせは該異なる出力状態を分離することで貧弱な仕事をしているので該適応度関数は低い値となる。代わりに、該適応度関数は、該部分空間内の個別セルの情報豊富さを調べ、次いで該セルの加重平均を形成することにより該部分空間の情報豊富さのメザーを提供してもよい。
【００５５】
好ましくは、出力状態クラスタリングのグローバルなメザーは最良のビット記号列の発展をドライブする該適応度関数として使用される。このメザーは好ましくはクラスタリングを規定する強力な方法であるエントロピー関数に基づくのがよい。適応度関数のこのエントロピー的規定を用いて、該出力を最も良くクラスターし分離する入力組み合わせを表すビット記号列が該発展型過程から出現する。代わりの適応度関数は、出力状態確率の標準偏差か分散か、又は少なくとも１つの出力確率が他の出力確率より顕著に大きい部分空間内のセル数を表す値かを含む。出力状態の集中を測定する他の同様な発見的方法（heuristics）、又はアドホック（ad hoc）な規則は発展型過程内で容易に交換される。
ｃ．発展型過程の詳細
１．Ｎビット２進記号列のランダムなプールの創生
図５Ａを参照すると、該発展型過程５００は過程５１０で始まり、そこではＮビットの２進記号列のランダムなプールが創られる。これらの初期２進記号列は、それらがともかく最適であると云う先験的理由がないので一般的にそれらの適応度関数用には非常に低い値しか持たない入力フイーチャー組み合わせをエンコードする。この初期プールは該発展型過程を始動するため使われる。
【００５６】
２．適応度の計算
該プール内の各２進記号列の適応度は過程（ｂ）で説明した方法を使用して計算される。該データは過程５２０で示すようにバランスを取られる。各２進記号列用にフイーチャー部分空間が発生され、データベース内のデータが対応する部分空間内へ射影される。該部分空間は過程５３０で行われた選択に従って、等間隔のビニング５３２又は適合的に隔てられたビニング５３４の選択に依りビンに分けられる。考慮下の特定の遺伝子が過程５４０で選択され、そしてビンの数は過程５５０で、好ましくはユーザー入力により、ビンの固定数５５２を指定するか又はセル当たりサンプルの平均数５５４を指定することにより決定される。該ビン配置は次いで過程５６０に示す様に、決定される。次いで対応２進記号列の適応度を表す出力状態のクラスタリングと分離の程度を計算するためにエントロピー関数又は他の規則が使用される。これは、データ点が各部分空間内に配置される過程５７０と、グローバル情報コンテントが決定される過程５８０で示される。過程５８５により示される様に、次の遺伝子シーケンスは過程５４０の開始で動作する。
【００５７】

３．適応度の加重ルーレットホイール（weighted rourette wheel）の創生
各２進記号列の適応度が計算された後、該適応度の加重ルーレットホイール５９２が図５Ｃに示す様に創られる。これは、より高い適応度値（fitness value）を有する２進記号列がより低い適応度値を有する２進記号列よりも比例してより広いスロット幅に付随される過程と考えられる。これは、該ルーレットホイールが廻されると、より低い適応度の２進記号列よりも、より高い適応度の２進記号列の選択に、より重く加重する。この過程は下記で更に詳細に説明する。

【００５８】

４．新しい親の２進記号列（new parent binary strings）の選択
ルーレットホイール５９２は次いで廻され、該ホイールが終わるスロットに対応する２進記号列が選択される。もし元のプールにＮ個の２進記号列があるなら、該ホイール５９２はＮ個の新親記号列を選択するためＮ回廻される。ここで重要な点はもしそれが高い適応度値を有するなら該同じ２進記号列が１回より多く選ばれ得ることである。逆に、低い適応度関数を有する２進記号列は、それが完全に排除されることはないが、親として決して選択されないことが起こり得る。次いでＮ個の親が、新しい子の２進記号列発生への先駆者としてＮ／２個の対に対化される。

【００５９】

５．子記号列を創る親の交叉（crossover）と突然変異（mutation）
一旦２つの親が選ばれると、図５Ｄに示す、交叉オペレーション（crossover operation）５９４が行われるべきか否かを決定するために加重コインがフリップされる。もしこれが交叉オペレーションとなるなら、クロシングサイトがビット位置１と該記号列内の最後のビット位置の次にあるの最後の起こり得るクロシングサイトとの間でランダムに選択される。該クロシングサイトは各親を右側と左側に分割する。図５Ｄに示す様に、各親の左側を他の親に右側と連結することにより２つの子記号列が創られるが、そこでは該親遺伝子１０００１と０００１１は左半分１００と０００、そして右半分０１と１１に分割され、次いで１００１１と０００１１を形成するよう組み合わされる。最後に、該２つの子記号列が創られた後、該子記号列プールの多様性を増やすために該子記号列の小数の個別ビットがランダムに逆にされる（突然変異される）。これは与えられたビットが逆にされる確率に換算して指定出来る。逆転の確率は望ましいビット突然変異の数と該記号列内ビット数に基づいて尺度合わせされる。すなわち、もし記号列当たり平均５つの突然変異が望まれるならば、与えられたビット変更の確率は１００ビット記号列用に０．０５に、そして５０ビット記号列用に０．１等に設定される。

【００６０】

６．発展型過程の継続
過程５９０に示す様に、上記過程２－５は、各創られた子記号列プールを次世代用の新しい親プールとして使用して、数回（又は数世代）繰り返される。該子記号列プールが発展すると、それらの対応適応度は平均で改善すべきであるが、それは各世代で、新しい子記号列を創るために、より適した記号列が優先的にメートされるからである。

【００６１】

該発展型過程は、予め決められた数の世代の後か、又は最高適応度の記号列か又は平均プール適応度か何れかが最早変化しない時か、何れかで停止出来る。

【００６２】

最適化問題を解くための遺伝的アルゴリズムの使用で、解かれる必要にある２つの重要な項目がある。第１の項目はエンコーデイングスキームである。該問題がビット記号列としてエンコードされ得る解の役に立つか？第２の項目は該適応度関数の選出である。該発展型過程は該適応度関数により統制される（すなわち、導かれる）ので、その解の質は間近な目標への適応度関数のマッチングに密接に依存している。

【００６３】

ここに説明した好ましい方法では、第１の項目は、図４で図解され、各ビットがデータ集合のＮの入力の１つと対応する、Ｎビット２進フイーチャービット記号列を含む遺伝子を

規定することにより解決される。該Ｎビット２進フイーチャービット記号列の各ビットは対応入力を参照し、もし該対応入力が該フイーチャー部分空間内にあれば該値１を、もし該対応入力が該フイーチャー部分空間内に無ければ該値０を有する。

【００６４】

該好ましい方法では、第２項目はフイーチャー部分空間のグローバルエントロピーを計算する情報エントロピーメザー（informational entropy measures）を使用することにより解決される。該フイーチャー部分空間のグローバルエントロピーは、それから最適モデルが発展させられ得る最適フイーチャー組み合わせのプールの発展をドライブする適応度関数として使用される。該グローバルエントロピーは、フイーチャー部分空間内のセルのローカルエントロピーを最初に決定し、そして該ローカルエントロピーの加重和として全体のフイーチャー部分空間のグローバルエントロピーを計算することにより計算される。代わりに、部分空間のグローバルエントロピーは、該全体の部分空間の間で、与えられる出力用の点の分布を調べ、そして次いで全ての状態に亘り特定状態向けエントロピーの加重平均を形成することにより決定されてもよい。フイーチャー部分空間プールを保持する能力は、そのどちらも最終モデルのローバストさに寄与する該解空間内の冗長度と多様性の双方を提供する。

ローカルセルエントロピーとグローバル部分空間エントロピーの決定

好ましい方法の側面に依れば、情報コンテントのレベルが測定される。特に、セル又は部分空間の情報コンテントのレベルはデータ分布の均一性のメザーである。すなわち、データが均一である程、システムのモデル化の目的にそれが持つ予測価値は大きくなり、従って、情報コンテントのレベルは高くなる。該均一性は多数の代替え的方法で測定されてもよい。１つのこの様な方法はクラスタリングパラメーター（clustering parameter）を使用する。用語クラスタリングパラメーターはローカルセルエントロピー、考慮下の特定部分空間上で計算された特定出力のエントロピー、又はここで論じられる発見的方法、又は他の同様な方法を指す。

【００６５】

図６を参照すると、個別セルの情報コンテントは方法６００により示されたカテゴリー的出力システム及び方法６０２による連続する定量的モデル用に決定される。好ましい実施例では、前に論じたニシ（Nishi）の情報エントロピー規定が、該情報コンテントを表すローカル及びグローバル両エントロピー加重を数学的に規定するため使用される。本発明の実験型モデリング用には、ニシにより拡張された、シャノンのエントロピーの概念が、該エントロピーのメザー（measure）が計算されるデータ集合用の適当なメザーであることが見出されて来た。ニシの式が出力状態に対応する確率の集合に適用される。等しい出力確率を有するセル（各出力が等しく似ている）は少しの情報コンテントしか有しない。かくして、高い情報コンテントを有するデータ集合は他より高い、幾らかの確率を有する。より大きな確率的変動（greater probabilistic variations）は出力状態の不平衡（imbalance in the output states）を反映し、従って該データ集合の高い情報豊富さの指標を与える。

【００６６】

好ましい方法では、一般的なエントロピー加重項（general entropic weighting term）Ｗが規定され、Ｗ＝１－Ｅの形式を有する。該エントロピー加重項Ｗはニシの情報エントロピー関数Ｅの補数（complement）であり、完全に不均一な分布用に値１を有し、完全に均一な分布用に値０を有する。

【００６７】

図６の方法６００を再び参照すると、情報レベルはローカルエントロピー加重項（local entropic weighting term）を計算することにより決定される。例えば、部分空間内の与えられたセル用に適当なものは次の仕方で規定され得るが、すなわち最初に、過程６１０で、n_Cエントリーを有するデータ集合が創られ、ここでn_Cは出力状態の数である。各エントリーは下記で与えられるセルｉ用の特定状態向けローカル確率$p_{C|i}$に対応しており、

【００６８】
【数５】

$$p_{c|i} = n_{ci} / \sum_{k=1}^{n} n_{ki},$$

【００６９】
ここでｎ$_{Ci}$はｃの出力状態を有するセルｉ内の点の数であり、該和はセルｉ内の全ての出力状態ｋに亘り延び、かくしてセルｉ内の全ての点を含む。与えられセルｉ用に、値ｐ$_{C|i}$のシーケンスは種々の出力状態ｃにある確率を表す。過程６２０で該セルの情報コンテントは決定される。好ましくは、ニシの情報エントロピー規定が部分空間Ｓ内の与えられたセルｉ用のローカルエントロピー項Ｅを規定するため使用されるのがよく、
【００７０】
【数６】

$$E_i^S = (\sum_{k=1}^{n_c} f_{k|i}^S \ln f_{k|i}^S) / \ln(1/n),$$

【００７１】
ここで和の変数ｋは出力状態、ｎ$_{C}$は出力状態（又は”カテゴリー”）の総数を表し、そして
【００７２】
【数７】

$$f_{k|i}^S = p_{c|i}^S / \sum_{k=1}^{n_c} p_{k|i}^S.$$

【００７３】
である。
【００７４】
勿論、全てのｋに亘る全てのｐ$_{k|i}$の和は１に等しいが、明確化のため上記に含まれる。
【００７５】
最後に、又過程６２０で、該ローカルエントロピー加重係数は
Ｗ$_{i}^{Ls}$＝１－Ｅ$_{i}^{s}$
であり、ここで上書きＬｓはWが部分空間Ｓ内でセル用のローカルエントロピー関数であることを呼称する。高い情報コンテントを有するセルは高いローカルエントロピー加重を有する。すなわち、それらはＷ$_{i}^{Ls}$の高い値を有する。
【００７６】
代わりに、該情報コンテントは、該出力確率値の分散又は標準偏差を決定することによるか、又は何等かの１つの出力が予め規定されたしきい値を上回る付随確率を有するかどうかを決定することによる様な、均一性のもう１つのメザーにより測定されてもよい。例えば、セルの確率分布に基づきセルに値を割り当ててもよい。特に、予め決められた値より大きい何等かの出力状態確率を有するセルは１の値を割り当てられ、該出力状態確率のどれも予め決められた値より大きくないどのセルも値０を割り当てられる。該予め決められた値は該フイーチャー部分空間（モデル、フレームワーク、スーパーフレームワーク等）の結果に基づき実験的に選ばれた定数である。該定数は又出力状態の数に基づいてもよい。例えば、何れかの出力状態が平均より大きい発生の尤度（greater-than-average likelihood of occurring）を有するセルの数を数えたいと願ってもよい。それで、ｎの出力状態システムについて、１／ｎより大きい何等か１つの出力状態確率を有するどんなセルも１の値を与えられるか、又はｋ／ｎより大きければ、或る定数ｋが与えられる。他のセル

はゼロの値を与えられる。

【００７７】

代わりに、セルに与えられる加重は与えられた確率を越える出力状態の数に基づいて増加出来る。例えば、４出力状態システムでは、０．２５より大きい発生確率を有する２つの出力状態を有するセルは２の加重を与えられる。更に進んだ代替えとして、セルの又はグローバルな加重は出力状態の分散に基づくことが出来る。他の同様な発見的方法が考慮下のセルの情報コンテントを決定するため使用されてもよい。

【００７８】

モデル化されつつある過程の出力が連続的な場合、ローカルエントロピーは方法６０２に示す様に計算される。過程６３０で、該セルに存在する出力値の全てを含むデータ集合が創られる。該セルの情報コンテントは過程６４０で計算される。出力に特定的な確率を処理する時、高い情報コンテントを有するデータ集合は他より高い或る確率を有することが思い出される。出力値を直接処理する時、しかしながら、過程６３０－６７０でその場合である様に、情報豊富な集合はより均一なデータ値を有するそれらである。すなわち、高い情報集合は出力値ではより少ない変動を有する。かくして、もし情報コンテントが該ニシのエントロピー計算を使用して決定されれば、該補数的値１－Ｅを形成する必要はない。この場合の加重係数は簡単にニシのエントロピーＥに等しい。

【００７９】

加えて、過程６５０と６６０で示す様に、低エントロピーセルにゼロを設定するようにしきい値限定を適用することが望ましい。これはグローバルな計算が行われる時意味のない情報コンテントを有するセルの情報コンテントを累積することに付随する誤った影響を制限する助けになる。ローカルなセルのエントロピーの計算は過程６７０に示す様に完了する。

【００８０】

代わりに、連続的出力システムを取り扱う時、該出力を複数のカテゴリーに量子化し、各量子化レベルでの確率を有するデータ集合を規定するために、過程６１０で示す上記方法の過程を使用することが可能である。残りの過程６２０も、上記説明の様にエントロピー加重を計算することによって、該情報コンテントを決定するため行われる。

ローカルエントロピーの加重和としてのグローバルエントロピーの計算

図７を参照すると、部分空間Ｓ用のグローバルエントロピーW^{gs}は次いで、その部分空間内の全セルに亘りローカルセルエントロピーW^{ls}のセル母集団加重和（cell-population-weighted sum）として計算される。

【００８１】

【数８】

$$W^{gs} = \sum_{i=1}^{n} n_i^s W_i^{ls} / \sum_{i=1}^{n} n_i^s ,$$

【００８２】

ここでｎは部分空間Ｓ内のセル数を表し、n_i^sは部分空間Ｓ内のセルｉ内のカウント（データ点）数を表す。実際は、これは、それがその部分空間内のセルのピューリテイ（purity）の全体的メザーを記述するので、グローバルエントロピーの有用なメザーであることになった。図８はローカルとグローバルの情報コンテントの計算を図解する。図９はローカルとグローバルのエントロピーパラメーターの例を示す。高い情報コンテントを有する部分空間はW^{gs}の高い値を有する。出力状態依存のグローバルエントロピーを計算する代替え的方法

規定された基本的統計量は、該出力が部分空間Ｓ内の状態ｃ内にあるとした場合にセルｉ内にある確率を表す確率$p_i|_c$である。

【００８３】

【数９】

$$p_{ikc}^{S} = n_{ci} / \sum_{j=1}^{n} n_{cj} ,$$

【００８４】
ここでｎ$_{ci}$は出力状態ｃを有するセルｉ内の点の数であり、該和は部分空間Ｓ内の全てのセルｊに亘って伸展する。
【００８５】
該ニシの情報エントロピー規定が部分空間Ｓ内の与えられた出力状態ｃについてグローバルエントロピー項Ｗ^{gs}$_{c}$を規定するため使用出来る。最初に、与えられた状態ｃ用のニシのエントロピーが計算される：
【００８６】
【数１０】

$$E_c^s = (\sum_{i=1}^{n} f_{ikc}^{s} \ln f_{ikc}^{s}) / \ln(1/n)$$

【００８７】
ここでｎはセル数であり、
【００８８】
【数１１】

$$f_{ikc}^{s} = p_{ikc}^{s} / \sum_{j=1}^{n} p_{jkc}^{s} .$$

【００８９】
である。
【００９０】
再び、状態に特定的な確率（state-specific probabilities）の全てのセルに亘る和である、分母は１に等しいが、一貫性と明確化のために上記表現に含まれる。Ｅ^{S}$_{C}$はかくして該部分空間Ｓ上の確率ｐ^{S}$_{i}$$|_{c}$の分布のグローバルな均一性を表す。最後に、該グローバルエントロピー項Ｗ$_{c}$^{gs}は下記で規定され
Ｗ$_{c}$^{gs}＝１－Ｅ^{S}$_{c}$
それは部分空間Ｓ内でのカテゴリーｃ用のグローバルな出力に特定的なエントロピー加重項である。これは、それが全体の部分空間を通しての点の分布（出力ｃに対応する）のクラスタリングを表す意味でグローバルなメザーである。高い情報コンテントを有する部分空間は高い値のＷ$_{c}$^{gS}を有する。
グローバルエントロピー加重係数の代替え的規定用のカテゴリーから独立した一般化
全カテゴリーに亘り加算することにより、代替え的グローバルエントロピー加重係数はカテゴリーから独立したグローバルエントロピー加重係数として規定され
【００９１】
【数１２】

$$\bar{E}^{S} = (\sum_{c=1}^{n_c} \sum_{i=1}^{n} f_{ikc}^{S} \ln f_{ikc}^{S}) / \ln(1/n')$$

【００９２】
ここでｎ’は＝ｎ$_{c}$ｎで、それは出力状態数とセル数の積であり、ここでは
【００９３】

【数１３】

$$f_{ijc}^{S} = p_{ijc}^{S} / (\sum_{c=1}^{n_c} \sum_{i=1}^{n} p_{ijc}^{S}).$$

【００９４】
である。勿論、上記式の分母は
【００９５】
【数１４】

$$\sum_{c=1}^{n_c} \sum_{i=1}^{n} p_{ijc}^{S} = n_c$$

【００９６】
と簡単化され、それはニシの式で使用される確率が適切に正規化されることを示す。この代替えの規定は出力状態数が多く、そして計算効率が望まれる状況で有用と信じられる。
【００９７】
上記議論で、該システムの出力値が離散的（discrete）、又は”カテゴリー的（categorical）”であることが仮定されている。同じ方法は、エントロピー計算の前に最初に出力値を離散的状態又はカテゴリーに人工的に量子化することにより、例え該出力値が連続的であっても、ローカル及びグローバルエントロピーを計算するため使用される。
【００９８】
トレーニングのデータ集合の出力状態の母集団の分布は該モデルの究極的有効性（ultimate validity）に付随されることは述べる価値がある。上記解析で、該データ集合はバランスされていると仮定されてもいるが、しかしながら、この様なことは常にはその場合ではない。２つの出力状態、ＡとＢとがある問題を考える。もし該トレーニングデータ集合が状態Ａを表すデータ項目から主として成るならば、該母集団の統計はアンバランスとなり、ことによると偏倚されたモデルの創生となる。インバランスの理由は、データコレクター（data collector）の部分での偏倚か、又は該データ集合の親母集団特性にある真性のインバランスか何れかである。
【００９９】
該データコレクターの部分での偏倚の場合、セル内の母集団統計がデータ項目の絶対数より寧ろ該セル内に存在する与えられた出力状態のデータ項目の部分を参照するように簡単な正規化が行われ得る。この正規化は多くの実験データ集合で成功裡に使われて来た。第２の場合では、該インバランスは”真実（real）”であるので、正規化は適当ではないかも知れない。
【０１００】
データ正規化の例は次の様である。
【０１０１】
２つの出力状態ＡとＢがある１００項目を有するデータ集合を考える。状態Ａに対応する７５項目と状態Ｂに対応する２５項目とがあると仮定する。状態Ａに対応する５項目と状態Ｂに対応する５項目を有する全部で１０項目がある部分空間内のセルを考える。絶対項では、我々は各エントリーが特定の状態用のカウントを参照する｛５，５｝に対応する”カウントデータ集合”を有するので、これはインピュアセル（impure cell）である。しかしながら、該データは次の様にその状態用の全体のカウントに対して各カウントを正規化することによりバランスさせられてもよい。
【０１０２】
【表１】

状　態	カウント	全体の分数
A	5	5／75＝1／15
B	5	5／25＝1／5

【0103】
該表からの該分数的カウントは次いでエントロピー計算で使用される。
【0104】
データ集合DはD＝｛1／15、1／5｝、d_{total}＝1／15＋1／5＝4／15を伴い、正規化されたデータ集合FはF＝｛1／4，3／4｝となる。エントロピーEは次の様に計算される。
【0105】
E＝｛0．25ln（0．25）＋0．75ln（0．75）｝／ln（1／2）＝0．811
変型されたニシのエントロピーWは1－E、すなわち1－0．811＝0．189である。図2Cはデータ集合内で与えられた出力状態が支配的な時データの影響をバランスさせる方法を図解するブロック図である。
予測指向の適応度関数を用いたモデル発展
一旦入力が量子化され、フイーチャー部分空間のプールが遺伝的アルゴリズムにより初めに同定されると、それらの好ましい部分空間の組み合わせを形成することによりモデルが発生される。上記説明の様に、データ又はトレーニングデータ集合と呼ばれるデータの部分集合は、そこから情報が抽出され得る多くのフイーチャー部分空間トポグラフイ（feature subspace topographies）を創るために使用される。高い情報コンテントを有する部分空間が一旦同定されると、これらの部分空間は、出力予測の目的で該データが内部へ射影される”ルックアップ（look up）”部分空間として使用される。
【0106】
特定の部分空間による出力予測は該特定の部分空間内の与えられたセル内の出力状態の分布により決定される。すなわち、各データ点（又はテストデータ部分空間内の各点）は、図3A－Cに関係して見られる様に、与えられた部分空間内の1つのセル内に分類される。各データ点に付随する出力を予測しようとして、人は、部分空間（全体のデータ集合、又はトレーニング部分集合）を占めるため使用されるデータの分布を単に見て、予測に到達するためこれを使用する。特定の部分空間による出力予測用に従う簡単な規則は、該出力が状態cにあるとなるべき確率が$p_{c|i}$により与えられることである。この”ローカル”確率はフイーチャー部分空間内の与えられたセルを占めるサンプル点の出力分布を単に表している。
【0107】
与えられたモデルは部分空間の組み合わせであり、従って、該モデル内の考慮下の全ての部分空間に関して各点が調べられる。該ローカル確率は本質的に”ベース（base）”量であり、それは次いでモデル内のローカル及びグローバルの両エントロピーにより加重される。該用語”ローカルエントロピー”と”グローバルエントロピー”は”エントロピー的係数”又は”エントロピー的加重”としてここでは集合的に引用される。それは、簡単な確率的モデルと比較した時本方法をかなりより精密化するモデル予測を決定するグローバル及びローカルの両方の情報定量評価（information metrics）の追加である。このエントロピー係数の目的は”情報豊富”な部分空間内の”情報豊富”なセルを際立たせ（emphasize）、個別的に情報が貧弱か｛すなわち、情報豊富さの少ない（less information-rich）｝、又は情報貧弱な部分空間内に置かれるか何れかであるセルを軽視（de-emphasize）することである。

【０１０８】
かくして発展型モデル過程をドライブするため使用される各部分空間組み合わせ又はモデル用の適応度関数は、予測のエントロピー的加重和と、該予測と該テストデータ点に付随する実際の出力値との間の付随誤差率（associated error rate）とである（再び、全体データ集合か又は部分集合かの何れか）。
【０１０９】
かくして、該方法の１側面に依ると、ローカル及びグローバルエントロピー加重係数は該フイーチャー部分空間の情報コンテントを特徴付けるために使用される。フイーチャー部分空間セルの寄与をローカル及びグローバルな情報メザーにより加重することにより、該方法は種々の種類のノイズ源を有効に抑制することが出来る。１つのこの様なノイズ源はセル内のローカルノイズである。もしセル内の出力状態の分布が均一であるなら、そのセルは少しの予測情報しか有しない。与えられた出力状態の確率はセル内の出力状態の全分布の性質をほのめかすことは出来るが、それは全体の物語は述べない。全ての他の出力状態の分布は与えられた出力状態の確率内には含まれない。２進出力システムの他の何れでも、１つの出力状態確率内に含まれた情報はかくして不完全である。個別セルに付随するローカルエントロピー項の計算は全体のローカル確率分布を特徴付ける加重係数となる。
【０１１０】
上記説明の様に、該グローバルエントロピー係数は比較目的に幾つかの異なる方法で計算出来る。部分空間のグローバルエントロピーを規定する好ましい技術はグローバルエントロピーをローカルセルエントロピーのセル母集団加重和（cell-population-weighted sum）として規定することである。該ローカルエントロピーは部分空間内の各セル用に計算され、この部分空間用の該グローバルエントロピーは次いで全てのセルに亘りセル母集団加重和を行うことにより計算される。これは部分空間について全体のグローバルセル情報エントロピーを測定する（部分空間のセル全部上で）。
【０１１１】
代わりのグローバルメザーは全体の部分空間上で該セル内の各出力状態の確率分布を調べる。もしこの分布が均一なら、関心のある該部分空間はその出力状態について少しの予測情報しか有さない。この実施例で、部分空間内で各出力状態用に別々のグローバルエントロピー項が計算される。この代わりのグローバルエントロピー項は、各出力状態用に同じである、前に説明したグローバルエントロピー項とは異なる。この代わりのグローバルエントロピーのメザーは、与えられた部分空間が１つの出力状態に関しては”情報豊富”であるが、異なる出力状態に関しては”情報が貧弱”である可能性を受け入れる。
【０１１２】
本方法はノイズを抑制するためにローカル及びグローバルの両方のベースの加重係数の独立した計算を考慮する。これらの係数は最大の予測精度用にローカル及びグローバル情報の間の最適バランスを得るために個別に調整、又は”ツイーク（tweaked）”される。多くの従来技術のデータモデリングシステムでは、ローカル及びグローバル加重係数の相対的大きさを便利に調整することは難しい。前記の様に、大抵の従来技術の方法は解に到達するために全体のデータ集合上での目的関数（objective function）の最適化に依存する。
【０１１３】
もう１つの関連項目は冗長度（redundancy）のそれである。幾つかの入力フイーチャーは与えられた出力に関する本質的に同じ情報コンテントを含んでいる。例え２つのフイーチャーが特定の出力状態に関する情報を含まなくても、それらはなお相関しているかも知れない。冗長度は本発明の方法を本質的に制限せず、事実、それは全体の計算コストを増やすけれども、創られるローバストさを該モデルに組み入れる方法として非常に役立ち得る。情報メザーを使用するクラスタリング方法はフイーチャー間の冗長度を同定するために利用可能であり、下記で論じる。
【０１１４】
ローカル及びグローバルの両方のエントロピー加重係数は分布の”構造”量（amount of

"structure"）を測定する。分布がより少ししか均一でない、又は”より多く構造化されて（more structured）”いる程、その対応するエントロピー加重Wはより高い。データ空間の構造のこの側面はローカル及びグローバルの統計の重要性を加重するため使用される。

【０１１５】

ローカル及びグローバルの両エントロピー項の計算は該方法でのローカル及びグローバルな情報加重係数の別々な制御を考慮する。生ずる自然な問題はローカルさの規定であり、ローカルとはどれ程ローカルなのか？この質問の回答は勿論取り組まれる特定の問題による。好ましい実施例に依れば、該方法は該ビンの解像を走査することによりローカルさの最良の説明をシステム的に探索するが、該解像度は今度は最高の予測精度を提供するために多次元のセルサイズを決定する。特に、情報豊富なフイーチャー部分空間の異なるグループが同定され（エグゾースチブな探索か又はフイーチャー部分空間発展かの何れかにより）、そこでは各グループは部分空間当たり異なる数のセルｎを使用する。事実、セル数ｎは最小値から最大値までエグゾースチブに探索される。セルの最大数はセル当たりの点の最小平均の意味で指定されるが、それは余りに多くのビンで部分空間の分解能を上げ過ぎることは望ましくないからである。最小数は１より例え小さくてもよい。

【０１１６】

この点で出力状態の特性をより詳細に考慮することは余談に入る価値がある。本発明の方法では、入力の量子化は多次元部分空間を創るために行われる。分類問題では、該出力変数は離散的カテゴリー又は状態であり、かくして既に量子化されている。定量的モデリングでは、出力変数は連続的である。この様な場合、１つの起こり得る解は該出力状態空間の離散ビンへの人工的な量子化を行うことである。該出力データ空間が量子化された後、上記で説明した離散的モデリングフレームワークがローカル及びグローバルエントロピー係数を測定するために使用され得る。これらのエントロピー係数は下記説明の方法を用いて該出力の連続値の予測に使用され得る。

【０１１７】

精度に関する重要なメザーは出力状態カテゴリーの数、n_cの平均全セル母集団統計に対する比$<n_{pop}>$である。もしn_cが$<n_{pop}>$より遙かに大きければ、大抵の出力状態はセル内で空いており、貧弱な統計となり、モデルでの起こり得る劣化となる。これは再びより多くのデータを主張し（argues for）、それはデータドライブされるモデルには当然である。コンピュータハードウエア技術の進歩と共に、多量のデータ集合の取得と記憶の能力は急激に増加し、本発明の方法は該データからの情報抽出を可能にする。該方法は、n_cの値が小さい（１－１０の桁で）多くの真実の世界の問題でn_cが$<n_{pop}>$より遙かに大きい時でも驚く程良く作動することが分かった。これは多数の部分空間上での加算統計の協力効果のためかも知れない。

【０１１８】

抄録すると、フイーチャー部分空間に付随するグローバルエントロピー係数は、遺伝的アルゴリズムを使用して最も情報豊富なフイーチャーのプールを発展させるため使用される適応度関数として使用され得る。このプールの決定は前に説明したデータ量子化条件に依存する。セル当たりサンプル点の平均数が減少すると、該ローカル及びグローバルエントロピー情報メザーは一般に増加する。しかしながら、これは、これらの量子化条件が最終モデルの開発で良く一般化することを必ずしも意味しない。実際に、セル当たりサンプル点の平均数が１より可成り少ない（すなわち、０．１以下）量子化条件下でフイーチャーを発展させることはなお精確なモデルに帰着する。これは主に、該フイーチャープール内の多数の部分空間上での加算統計の協力効果のためである。

システム入力からシステム出力を最も精密に予測するフイーチャーデータ集合の部分集合の決定

図１０を参照すると、高い情報エントロピーを有するフイーチャーデータ集合が一旦決定されると、このフイーチャー集合は予測モデルを直接開発するため使用されてもよい。しかしながら、発展型方法（evolutionary method）を使用する該フイーチャー選択過程は

、比較的高い情報エントロピーを有する高次元数データ空間内でそれらのフィーチャーのみを保持することによりいわゆる”次元数の災い（curse of dimensionality）”を緩和する可成りの利点を有する。この関係で、Ｎ次元空間内の起こり得る２進フィーチャービット記号列の総数は2^Nであり、その量はＮと共に指数関数的に増加することを注意すべきである。

【０１１９】

一旦フィーチャーデータ集合が決定されると、どんなサンプルデータ点用にも出力状態確率ベクトルを計算することが出来る。図１４を参照すると、このベクトルを計算するためには、全加重係数を創るよう該ローカル及びグローバルエントロピー加重係数を組み合わせることが最初に必要である。本発明の方法では、該ローカル及びグローバルエントロピー加重を含む一般的第３次表現が最適モデル性能用に実験的に調整された係数を用いて規定される。該全加重係数用の一般的表現はかくして次の様に見られる。

【０１２０】

$W^S{}_{ic}=a(W^{ls}{}_i)^2W^{gs}{}_c+b(W^{gs}{}_c)^2W^{ls}{}_i+c(W^{ls}{}_i)^2+d(W^{gs}{}_c)^2+eW^{ls}{}_iW^{gs}{}_c+fW^{ls}{}_i+gW^{gs}{}_c+h$

かくして、各部分空間Ｓ内の各セルｉは該与えられた部分空間Ｓ用の該ローカル及びグローバル加重の組み合わせである付随する一般的加重係数W^Sを有する（該式は又グローバル加重係数Ｗｇｓが出力状態依存性であり、従って該一般的加重係数が出力状態依存性であることを示すことに注意を要す。該グローバル加重係数が全ての出力状態に亘って計算される場合、出力状態ｃへの依存は除かれる）。

【０１２１】

ａからｈまでのパラメーターは最も精密なモデル、フレーム、スーパーフレーム他を得るために実験的に調整される。多くの問題では、該グローバルエントロピー回数も存在するが、該加重係数は該ローカルエントロピー加重係数により支配される。それはここで説明される方法がフィーチャー部分空間内のローカル統計に可成りの重要性を提供する点を強化し、それはここに説明される方法と従来技術のモデル化の取り組みとの間を際立たせる特徴である。該モデル用の信頼限界の確立の中では、該モデル係数は該誤差統計を計算するために変更され得る。

【０１２２】

一旦$W^S{}_{ic}$用の適当な値が決定されると、サンプル点ｄ用の各出力状態の確率は次の様に計算出来る。

【０１２３】

【数１５】

$$P_c(d)=\sum_{s=1}^{n_s}W^s_{i_dc}p^s_{ci_d},$$

【０１２４】

ここで該加算は全n_s部分空間上に延び、サンプル点ｄは各部分空間内の対応するセルi_d内へ射影するよう仮定され、該ローカル確率$p_c|_{id}$は該点がセルi_d内へ写像する事実がある時、該出力が状態ｃである確率である。上記の様に、もし一般的エントロピー加重が出力依存でないならば、一般的エントロピー加重の下付き文字ｃは上記式で無視されてもよい。各出力状態ｃ用確率は次いで確率ベクトル内に組み合わされ得る。

【０１２５】

$P(d)=\{P_1(d),\ldots,P_{Kc}(d)\}/N(i)$

ここでK_c出力状態が仮定され、そして

$N(i)=\Sigma P_c(i)$

は正規化係数で、確率の和が１であることを保証するために、ｃ＝１からK_cまでに亘り加算される。

【０１２６】

出力状態確率ベクトルＰ（ｉ）はサンプル点ｄの分類までの該データ空間内に含まれた情報を要約している。ニューラルネットワークの様な種々の従来技術のモデル化の取り組みも同様なベクトルとなり、異なる取り組みは該結果を解釈すると取られた。１９９４年発行の、レビューオブサイエンテイフイックインスツルメント（Review of Scientific Instruments）、６５巻（６）、１８０３－１８３２ｐｐ、ビショップ、シー．エム．（Bishop,C.M.）著”ニューラルネットワークとそれらの応用（Neural networks and Their Applications）”で説明される様に、共通に使用される方法は、予測された出力状態を発生の最も大きな確率を有する状態として割り当てる”勝者１人占め（winner take all）”戦術を使用することである。

フイーチャー部分空間の部分集合を使用する最適モデルの発展

高いグローバルエントロピー加重を有する部分空間を同定するための発展型方法は上記で論じられた。これは次元数の災い（curse）が明らかな多くの入力フイーチャーを有する問題で特に有用である。第１の発展段階では、該発展をドライブする適応度関数は部分空間のグローバルエントロピーである。最も良く予測するモデルを決定するために発展の概念を使うことも可能である。第２の発展段階では目標はテストデータ集合で最低誤差となる高いグローバルエントロピーを有するフイーチャー部分空間の最適部分集合を同定することである。この第２の発展段階は最良の予測モデルを作るために協力的仕方で”一緒に良く作用する（work well together）”部分空間をグループ化する。同時に該モデリング過程で追加的ノイズを導入する部分空間は第２発展段階中に間引かれる（culled）。図１５を参照すると、この第２発展段階での該適応度関数は次いで、フイーチャー部分空間の特定の部分集合を使用することから得られるテスト集合内の全体の予測誤差である。

【０１２７】

Ｍが予め決められている第１発展段階の後にＭのフイーチャーが高グローバルエントロピーを有するフイーチャー部分空間の最後の遺伝子プール内に存在すれば、フイーチャーの最適組み合わせを見出すために第２発展過程が使用される。Ｍビットの”モデルベクトル”が規定されるが、そこでは各ビット位置は与えられたフイーチャーの在り、無しをエンコードする。該モデルベクトルによりエンコードされた該フイーチャーを使用してトレーニングとテステイングが行われ、該適応度関数はテスト集合上のモデリング過程から生じる適当な性能定量評価である。分類問題用には、該適当な性能定量評価は該テスト集合内に正しく分類されるサンプルのパーセントである。定量的モデリング問題用には、該適当な性能定量評価は該テスト集合内の予測と実際の値の間の正規化された絶対差であり下記で与えられ

【０１２８】

【数１６】

$$F = 1 - \frac{\sum_{d=1}^{N} \frac{|a_d - p_d|}{d_{max} - d_{min}}}{N},$$

【０１２９】

ここでa_iはテスト点ｄ用の実際出力値、p_dは該テスト点ｄ用の予測値、d_{max}はテスト点値の出力範囲の最大値、そしてd_{min}はテスト点値の該範囲の最小出力値である。

【０１３０】

一旦第２発展過程が終了すると、最適モデルベクトルが該モデリング過程用の最適フイーチャー組み合わせを選択するため使用される。それで、第１発展段階は高情報エントロピーのフイーチャーのプールを同定したが、該プールはテスト集合内の予測誤差を最小にする最良部分集合のフイーチャーを見出すために該第２発展段階で更に発展させられる。この全体の過程は該モデリング問題への最良の実験的解を見出すために種々の発展的条件と制限下で繰り返される。

【０１３１】

かくして本発明の方法は階層的発展の概念を組み入れるが、そこでは最も情報豊富なフイーチャーのみならず、最良予測モデルを開発するために必要なフイーチャー部分空間の最適部分集合も、双方を同定するために、発展的方法が使用される。２つに発展段階を有することは該方法のユニークな利点を提供する。第１段階は手元の問題に見通しを得るために何れの次のモデリング過程からも独立して調べ得るフイーチャー部分空間の情報豊富な部分集合を作る。この見通しは今度は意志決定過程を導くため使用出来る。

【０１３２】

従来技術のモデリングパラダイムでの共通の苦言はそれらが入力フイーチャー内の何処に情報があるかを容易には明らかにしないことである。この欠点は従来技術の方法の能力を戦略計画と意志決定に参画することを制限する。本発明の方法では、第１発展段階の後の区切り点が、知的戦略計画と意志決定の可能性のみならず、次のモデリング過程が進める価値があるかどうかを決定する機会も考慮する。例えば、もし入力フイーチャーの充分豊富な集合が見出せないならば、本発明の方法は、ローバストなモデルを開発する前に、より情報豊富なフイーチャーを入力として含むデータへ戻るようモデル作成者（modeler）に指し示す。本方法はどの情報がないかを指定はしないが、本方法は充たされる必要のある情報ギャップがあることを指示する。情報ギャップ自体のこの指示は複雑な過程の理解で非常に価値がある。

情報写像の創生（Creation of Information Map）

図１１を参照すると、該第１発展段階の後、該問題の基本的理解を得るために該発展したフイーチャーデータ集合内に存在する入力の発生頻度のヒストグラムを作ることも又非常に有用である。このヒストグラムは該問題用の”情報写像（Information Map）”と規定出来る。幾つかの問題用には、該情報写像の構造は、入力の或る部分集合が入力の他の部分集合より可成り頻繁に起こるならば該問題の次元数を減らすために使用出来る。該部分集合の次元数を減らすことは、セル当たりサンプル点の平均数で部分空間を占めるために必要なデータ量が該次元数の増加につれて指数関数的に増加する様な次元数の災いのもう１つの側面を緩和する追加的利点を有する。図１２は遺伝子リストとその付随情報写像の例である。

エグゾースチブ（Exhausitve）な次元的モデリング

図１３を参照すると、もしこの様な次元数削減が可能なら、予測モデルは減少した入力データ集合を使用して開発可能である。本方法の好ましい実施例に依れば、Ｎの最も共通に起こる入力が該情報写像から同定され、次いでＮより小さいか等しい全てのＭ用に該Ｎのフイーチャーの Ｍの部分次元（sub-dimensions）内への全ての起こり得る射影（projection）が該フイーチャー部分空間を規定するため計算される。全てのこの様な射影を計算する帰納的アルゴリズム（recursive algorithm）は次の様である。

【０１３３】

フイーチャーの全ての組み合わせを計算する帰納的技術（recursive technique）は：各部分次元Ｍ用に、Ｎの数のリスト内で全てのＭケ組のもの（M-tuples）（長さＭの組み合わせ）を同定する問題を考える。第１要素が最初に選択され次いでＮ－１の数の残りのリスト内の全ての（Ｍ－１）ケ組のもの（長さＭ－１の組み合わせ）が帰納的仕方で同定される必要がある。一旦全てのこの様な（Ｍ－１）ケ組のものが同定され、該第１要素と組み合わされると、元のリストの第２要素が新しい第１要素として選択され、次いで該第２要素の過ぎた該Ｎ－２の残りの要素内の全ての（Ｍ－１）ケ組のものが同定される。この過程は該第１要素が該元のリストの終わりからのＭ＋１番目の要素を越えるまで続く。該アルゴリズムはそれがそれ自身を呼ぶので本質的に帰納的であり、それは又該要素の順序付けが重要でないことを仮定している。

【０１３４】

一旦与えられた部分次元Ｍ用の全てのフイーチャーの部分空間のプールが同定されると、このプールは、上記説明の方法を使用してテスト集合内の出力値を予測するために使用されるフイーチャー部分空間の集合として直接使用され得る。この過程は各部分次元Ｍ用の複数の量子化条件に亘って繰り返され得る。次いで最適な（部分次元、量子化）－対 {op

timum（sub-dimension, quantization）-pairs｝がテスト集合上の全予測誤差を最小化することに基づいて選択される。最適な（部分次元、量子化）対が選択された後、該最適な（部分次元、量子化）条件に対応するフイーチャー部分空間のプールは該第２の発展段階用のスタート点として使用され得る。この第２発展段階はテスト集合内に最小全予測誤差を有するこのプールからフイーチャー部分空間の最適部分集合を選択し、かくして最適モデルを規定する。

【０１３５】

一般的規則として、テスト集合上で充分な全予測精度をなお保存する比較的低い部分次元表現を決定することが有利と分かった。より低い部分次元で、より高いセル母集団統計が量子化の比較的精細なレベルに於いてさえもなお保持され得て、かくして該モデルの精度を改善する。

【０１３６】

もし元のデータ集合の次元が非常には高くないなら、エグゾースチブな次元モデリングの方法は元のデータ集合に直接適用され得る。これは高情報エントロピーを有するフイーチャーのプールを同定する第１発展過程を行う必要性を取り除く。

定量的モデリング

出力変数の人工的量子化を行うことによる定量的モデリング問題の分類問題への変換はローカル及びグローバルエントロピー係数を計算するために有用である。発生する自然な疑問は元のデータ集合内に存在する精度を如何に最終予測モデル内に保存するかである。これは、もし出力ビン解像度が乏しいセル統計を避けるためデータ集合のサイズにより抑制されるならば、特に重要である。伝統的分類問題用には、出力変数が起こり得る状態の離散的総体（ensemble）の１つを仮定出来るのみなので該精度問題（precision issue）は存在しない。

【０１３７】

出力変数の人工的量子化を行う１つの利点はローカル及びグローバル情報メザーの計算が、サンプル点の数から共に独立したカテゴリー又はセル上で加算が行われるシャノンの項に基づくことである。これはサンプル母集団統計を情報コンテントから分離することを容易化する。定量的モデリング用には、出力変数の人工的量子化は該ローカル及びグローバルエントロピーが同じ方法で計算されることを可能にして、かくしてサンプル母集団統計からの情報メザーの分離を保持する。

【０１３８】

出力変数量子化を使用してローカル及びグローバル情報メザーが計算された後、生の出力変数内の精度は最終予測モデル内の精度を回復するため使用され得る。

【０１３９】

最初に出力値の”スペクトラム”が全ての人工的出力変数カテゴリーに亘ってバランスを取られる。これは、各カテゴリー内の最終母集団が共通の目標値にあるように各出力カテゴリー内の各データ項目を或る尺度係数で有効に複製することにより達成される。典型的共通目標値はデータ点の全数を表す数である。

【０１４０】

データバランス化の１方法が上記で説明されたが、特定状態確率（state-specific probabilities）はその状態に対応する点の数に基づき正規化される。データを明確に複製することなくデータをバランス化する代わりの取り組みを下記で説明する。ニシの情報エントロピー項の計算は、Ｎがデータ集合のサイズを表す場合のｌｎ（１／Ｎ）係数を含む正規化項を有するが、この正規化は主にエントロピー項を０と１の間の値に制限するため役立っている。該正規化項は、均一性の程度が該データ集合のサイズに依存する問題に直接向けられていない。

【０１４１】

小さなデータ集合用には、該データ項目の該データ集合内の全データ項目の全体への正規化は微妙な偏倚を招く。例えデータ内の絶対的変動が比肩されるものでも、より小さいデータ集合内の正規化されたデータ項目間の相対変動は、より大きなデータ集合内の対応す

る項目間のそれより大きくなり得る。この偏倚を正すために、データバランス化過程が導入される。該バランス化過程を下記に説明する。

【０１４２】

２つのデータ集合D_1とD_2を考えるが、ここで該集合はそれぞれ、第１及び第２出力状態に対応する入力を表す。D_1はN_1項目を有し、D_2はN_2項目を有する。MがN_1とN_2の最小公倍数を、M_1とM_2が対応するデータ集合の各々用の掛け算尺度係数（multiplying scale factors）を表す。もしD_1をM_1倍、そしてD_2をM_2倍だけ複製するなら、最終両データ集合D'_1とD'_2はM項目を有する。必要な代数計算を行った後、新データ集合の各々用のニシのエントロピー項は次の様に変型される。

【０１４３】

$E'_1 = \{ln(1/M_1) + \Sigma f_i \ln f_i\} / \{ln(1/M_1) + ln(1/N_1)\}$

$E'_2 = \{ln(1/M_2) + \Sigma f'_i \ln f'_i\} / \{ln(1/M_2) + ln(1/N_2)\}$

ここでf_iとf'_iはそれぞれ元のデータ集合D_1とD_2上で正規化されたデータ部分を表す。

【０１４４】

もしセル内の出力データが密にクラスターされていれば、W_{local}は高い。逆に、もし該出力データが該セル内で全ての人工的出力カテゴリー上にばらまかれていれば、W_{local}は低い。該グローバルエントロピーは簡単に該部分空間内のセル上での数加重平均$<W^i_{local}>$として規定出来る。W_{global}は該部分空間内の情報の正規化総量を測定する。最後に、カテゴリーベースの分類で使用される基本確率定量評価P^s_{ic}は平均（又は代わりに中央値又は他の代表的統計量）セルアナログ出力値で置き換えられ得る。該部分空間上での平均セルアナログ出力値の加重和は次いで出力値を予測する離散的な場合に於ける様に行われることも出来る。それらの出力値で広いばらつき（spread）を有するセルは、個別セルが情報豊富でない部分空間でそうなる様に、下げて加重されることを注意する。

【０１４５】

セルの平均出力値μ^S_iの見積もりで、上記で規定したデータ複製尺度係数がバランス化されたデータ集合用にセル内平均値を計算するため使用される。該データバランス化過程はトレーニングデータ集合内の出力値の分布により導入される何等かの偏倚を除去するために行われる。

【０１４６】

【数１７】

$$\mu_i^s = \frac{\sum_{j=1}^{n} o_j M_j}{\sum_{j=1}^{n} M_j},$$

【０１４７】

ここでｎはセル内の項目の全数を表し、o_jは第ｊ番の項目の出力値を表しそしてM_jは第ｊ番のデータ項目に付随するデータ複製係数（data replication factor）を表すが、該データ複製係数は該第ｊ番の項目が属する人工的に量子化された状態に依存する。

【０１４８】

情報が貧者なセル及び部分空間からの”クリープ誤差（creep error）”を減らすために、オプションとして下記の過程が行われる。最初に、情報豊富な部分空間が離散出力状態の議論で前に説明した様に発展させられる。一旦最も情報豊富な部分空間が発展させられると、ローカル及びグローバル両エントロピーしきい値が、該情報豊富な部分空間に付随する平均値か又は中間値か何れかのエントロピー加重和の計算に向かって適用される。該

ローカルエントロピーしきい値より低いセル用ローカルエントロピー値はゼロ（０）に設定される。同様に、該平均の計算で誤差が徐々に累積されるのを避けるために、該グローバルエントロピーしきい値より低い部分空間用グローバルエントロピー値はゼロ（０）に設定される。

【０１４９】

該ローカル及びグローバルエントロピー関数のしきい値処理（thresholding）で、グローバルエントロピー関数の値の基づき該ローカルエントロピーの追加的しきい値処理を行うことが望ましいことが屡々ある。与えられた部分空間射影用のグローバルエントロピーがその対応するしきい値の下にあれば、その部分空間内の全てのセル用の該ローカルエントロピー関数はそれらの個別値に関係なくオプション的にゼロに設定出来る。前記説明のしきい値処理方法は又離散型出力状態モデリング用にもオプションとして行い得るが、クリープ誤差を最小化するためにより制限的過程が取られるべき定量的モデリング用でより高い価値がある。

【０１５０】

最後に、該しきい値処理過程を有しても有さなくても、本発明の方法はサンプルのテスト集合上で最小全出力誤差に帰着する情報豊富な部分空間の最適組み合わせを発展させ得る。又本発明の範囲内の定量的モデリングの方法は階層的発展をも含む。第１発展段階で、最も情報豊富な部分空間が、グローバルエントロピーを適応度関数として使用して、発展させられ、第２発展段階が続くがそこでは最小テスト誤差に帰着する情報豊富な部分空間の最適組み合わせが発展させられる。

【０１５１】

従来技術の方法に対する本発明の方法の利点はカテゴリー的及び定量的の両モデリングに共通のパラダイムが使用されることである。実験型のモデリングと過程理解とのための基礎としての分布状階層的発展の概念は、出力変数の唯１つ（連続型か離散型か何れか）の種類用にしか最適化されない従来技術の方法と対照的に、出力変数の両クラス（連続型及び離散型の両方）に適用される。

分布状階層的発展

ここに説明される方法は、”対象（object）”、例えば、フイーチャー、モデル、フレームワーク、そしてスーパーフレームワーク、の階層を創るために、情報理論からの概念を用いて、データの画像的表現、又はデータの多次元的表現の概念を使用する。用語”分布状階層的発展（distributed hierachial evolution）”は、モデル、フレームワーク、スーパーフレームワーク他の様な逐次より複雑で相互作用する発展型”対象”のグループが複雑なデータの漸進的により大きい量をモデル化し理解するため創られる発展型過程として規定される。大きな、複雑なデータ集合用には、前に説明したモデル創生過程が、最適モデルのグループを見出すために種々のトレーニング及びデータ集合上で繰り返される。最適モデルのグループの情報豊富な部分集合は次の様に決定される。

【０１５２】

図１６を参照すると、テストデータ集合の入力がモデルの選択された部分集合グループ（ランダムに選択されてよい）の各モデルに差し出され，各部分集合で予測される出力が各テストデータ出力と比較される。該部分集合で予測される出力の計算の過程は個別モデルを創るための過程と同様な仕方で行われ、そこでは個別のモデルで予測される値を入力としてそして実際の出力値を該出力として使用して、新しいトレーニング及びテストのデータ集合が創られる。この過程はモデルの多数の選択された部分集合グループ用に繰り返される。次いで該選択された部分集合グループは、”フレームワーク”と呼ばれるものを規定するためにシステム入力からシステム出力を最も精確に予測するモデルの最適部分集合グループを見出すために発展させられる。図１７Ａと１７Ｂはフレームワーク発展の概念を図解する。

【０１５３】

図１８Ａを参照すると、該フレームワーク創生過程は更に、最適フレームワークのグループを見出すためにモデル創生過程と同様な仕方で、繰り返される。最適フレームワークの

グループの情報豊富な部分集合は次の様に決定される。テストデータ集合の入力がフレームワークの選択された部分集合グループの各フレームワークに印加され、各フレームワーク部分集合で予測される出力が各テストデータ出力と比較される。フレームワーク部分集合で予測される出力を計算する過程は個別モデルを創る過程と同様な仕方で行われるが、そこでは新しいトレーニング及びテストのデータ集合が個別のフレームワークで予測された値を入力として、そして実際の出力値を該出力として使用して創られる。この過程はフレームワークの多数の選択された部分集合グループ用に繰り返される。該選択された部分集合グループは次いで、システム入力からシステム出力を最も精確に予測するフレームワークの最適部分集合グループ（これは”スーパーフレームワーク”と呼ばれる）を見出すために発展させられる。図１８Ｂはスーパーフレームワーク発展用の考慮を図解する。

【０１５４】

最適モデル決定過程、最適フレームワーク決定過程、或いは最適スーパーフレームワーク決定過程は、予め決められた停止条件が達成されるまで、繰り返されてもよい。該停止条件は、例えば、：１）予め決められた予測精度の達成、又は２）予測精度で更に進む改善が達成されない時、の様に規定されてもよい。本発明の方法はかくして実験データ集合上に分布した多数の相互作用する発展型対象の階層が同定される伸長可能な発展型過程である。発展対象の該階層の深さは解析されるべきデータ集合の複雑さにより決定される。簡単なデータ集合用には、全データ集合の非常に小さな部分集合を使用する１つのコンパクトなモデルで該全データ集合に亘りテストと検証（verification）のデータ集合値を精確に予測するのに充分である。該データ集合の複雑性が増加すると、該全データ集合（検証データ集合を含めて）を精確に説明するためにモデル、フレームワーク、スーパーフレームワークの階層を展開することが必要になるかも知れない。

【０１５５】

分布状階層的発展（Distributed Hierarchical Evolution）の顕著な計算的利点は、１つの大きな、モノリシックな実験型モデル（monolithic empirical model）の創生よりむしろ実験的モデルを規定するために大きなデータ集合に亘り分布された多数の、コンパクトな発展型対象の創生から生じる。高度に非線形の過程用には、大きなタスクを多くの小さいタスクに分けることが重要な実際的結果を有する顕著な計算的利点を提供する。

【０１５６】

分布状階層が成長すると、更に最適化が各段階で行われ、全体のデータ集合上での１つの、グローバル最適化上での顕著な性能改善となることは注意されるべきである。該大きなデータ集合内に含まれる益々増える情報は次々とより複雑な発展対象の相互作用の中に閉じ込められ、該相互作用は該実験型モデリング過程内の自由度の顕著な源として作用する。これは新データが現れた時該実験型モデルの更新を簡単化する。該実験型モデルの更新の初期過程は、該新データをテスト集合として使用して現在の実験型モデル内に最も最近の又は”最も高い”発展型対象の新グループを発展させることを含む。より早期のデータを使用して発展させられたより早期の又は”より低い”発展型対象は全く変えられる必要はないが該階層内の最も最近の発展型対象の新グループを創るため使用され得る。より早期の発展型対象のこのリクラスタリング（reclustering）からもし不充分に精確な新実験型モデルが生じるならば、その場合だけ、該新データの部分集合を使用して該階層内の該より早期の発展型対象を再発展（re-evolve）（該発展の繰り返し）させる必要がある。これが達成された時、最も最近の発展型対象の次ぎに新しいグループが該新データの異なる部分集合を使用して再発展させられる。モデル更新へのこのトップダウン的取り組みは、大抵の従来技術のモデリングの取り組みに共通なより伝統的なボトムアップのモデル更新に勝る顕著な計算的利点を供する。

監視されないフイーチャークラスタリング

部分集合用グローバルエントロピーメザーの概念は又入力相関に基づいてフイーチャークラスターを発展させるために適応度関数として使用される。例えフイーチャー部分集合内のセルが出力状態に関し可成りの情報を含まなくても、該セル母集団統計は該部分空間上でなお高度にクラスターされ得る。入力フイーチャー間の相関は、”グローバルエントロ

ピー加重係数の代替え的規定”の名称の節で前に説明したグローバルエントロピーパラメーターの代替えの規定と非常に似た情報エントロピー規定を使用して、出力状態から独立にセル母集団統計の均一性を計算することにより同定され得る。この場合、情報エントロピーを計算するために使用されたニシのデータ集合内の基本量はセル母集団であり、該ニシのデータ集合内のエントリーの数は該部分空間内のセルの数である。

【０１５７】

セル占有統計のグローバルエントロピーによりドライブされる発展型技術を使用して、最も高くクラスターされたフイーチャー部分空間は発展させられ、図１９Ａ、１９Ｂ、１９Ｃそして１９Ｄで示される。（１９Ａ及び１９Ｂの発展過程は図５Ａ及び５Ｂの前に説明した過程と同様である。考慮下の特定の遺伝子が過程７００で選択される。過程７４０により示す様に、次の遺伝子シーケンスは過程７００で始めに作動させられる。）

これは、クラスターを発見するための、１９９０年発行、アイイーイーイー論文集（Proceedings of the IEEE）７８巻４号１４６４－１４８０頁、コーネン、テー．（Kohnen, T.）著”自己組織化写像（The Self-Organizing Map）”で説明される様に、コーネンニューラルネットワーク（Kohnen neural networks）の様な他の監視されない方法の代替えである。この様な従来技術の方法に勝る本発明の方法の魅力的側面は監視されない及び監視されるモデリングの間の区別が、該エントロピー計算での出力状態情報の簡単な排除又は包含により非常に自然に起こることである。

【０１５８】

一旦高度にクラスターされたフイーチャー部分空間のプールが発展させられると、このプール内のフイーチャー部分空間のグループは、帰納用のドライブ条件としての該部分空間を横切る入力の重なり用に、例えば、しきい値条件を使用してより大きなクラスターを作るよう帰納的に合併させられ得る。この方法で、より大きなフイーチャークラスターのより小さなグループは、より大きなフイーチャークラスターの直接の同定が計算的に手に負えない非常に高い次元のデータ集合に於いても、効率良く同定され得る。

情報可視化

高いグローバル情報エントロピーのフイーチャーデータ集合を決定する第１の発展段階中に、該発展過程で同定される、最も高いローカル情報エントロピーを有するセルのリストを保持することも又可能である。

【０１５９】

乏しい、すなわち、人工的に情報豊富なセルのエントリーを避けるためにこのリストの選択では最小セルカウントしきい値が使用されてもよい。高いグローバル情報を有するフイーチャー内に存在するセルを調べることにより第１の発展段階の終わりでこの高いローカルエントロピーリストを創ることは可能である。計算効率の理由で、該第１発展段階の終わりでこの高いローカルエントロピーリストを創ることが好ましい。

【０１６０】

多次元データ空間内の情報豊富なセルを同定するこの方法は又”情報可視化（information visualization）”用にも使用出来る。多次元空間での情報可視化はデータ削減の問題として見られる。容易に理解可能な仕方でデータ集合内の本質的情報を取り込むために、最も情報豊富なセルのみが表示される必要がある。前の段落で、最も情報豊富なセルを選択するシステム的方法が論じられた。一旦これらのセルが全部分空間上で選択されると、カラー科学から得られた方法が視覚的に魅力ある仕方で該選択されたセルを表示するため使用されてもよい。例えば、カラー空間の｛色相（Hue）、彩度（Saturation）、明度（Lightness）｝特徴付けで、該色相座標が該セル出力カテゴリーへ写像され得る。該彩度座標はセルピューリテイ（cell purity）のメザーであるローカルセルエントロピー（$E^{Ls}{}_i$ か$W^{Ls}{}_i$の何れか）へ写像され得て、該明度座標は該セル内のデータ点の数（すなわち、該母集団）へ写像され得る。他の視覚的写像も行える。該第１発展段階の終わりでカテゴリー当たりのベースで最も情報豊富なセルのアクチブなリストを発生する過程は顕著なデータ減少過程に帰着したことは注意すべきである。このデータ減少は大きなデータ空間内で高い情報のローカル化された定義域（domain）の同定を容易にする。一旦全部分空間上

の走査が該第１発展段階の終わりで完了すると、このリストは適当な可視的写像方法を使用して適当な表示装置｛カラーシーアールテーモニター（color CRT monitor）の様な｝上に表示され得る。かくして多次元データ空間は表示目的で１次元リストへ減じられた。本発明の方法のユニークな側面は情報可視化に用いた方法論でデータモデリング行うため使用された方法論の組み合わせである。両方法用の共通した統合するカーネル（kernel）はセルと部分空間の形式でのデータの画像的表現を用いて情報エントロピーと発展を統合することにある。

ハイブリッドモデリング－分布状階層的発展のニューラルネットワーク又は他のモデリングパラダイムとの組み合わせ

本方法はデータモデリング用の強力なフレームワークを開示するが、どんなモデリングフレームワークも完全なものはないことを述べることは重要である。全てのモデリング方法は、その取り組み（approach）のためか又は該データに課された構造（geometries）のためか何れかで、”モデル偏倚（model bias）”を課す。分布状階層的発展はハイブリッドモデルを創るために他のモデリングパラダイムと組み合わされ得る。これらの他のパラダイムはニューラルネットワーク又は他の分類又はモデリングフレームワークであり得る。もし他の利用可能なモデリングツールが基本的に異なる哲学を有するなら、それらの１つ以上を分布状階層的発展と組み合わせることはモデル偏倚をスムーズ化する効果を有する。加えて、データ偏倚をスムーズ化するために種々のデータ集合を使用して多数の分散されたモデルが各パラダイム内に作られ得る。最後の予測結果は各モデルから来る個別予測の加重された又は加重されない組み合わせとなり得る。かくしてハイブリッドモデリングは、それが種々のモデリング哲学の強さを取り入れるので、極端に強力なフレームワークをモデリングに提供する。

法則の発見－分布状階層的発展の遺伝的プログラミングとの組み合わせ

第１発展段階の後、生じたフイーチャーデータ集合の情報コンテントを調べることは教示的（instructive）である。多くの場合、多数の比較的情報豊富なフイーチャーがあり、それは一緒に用いられると、実験型モデルの次ぎの展開用ベースを形成する。他方、もし、それらの絶対的情報コンテント（０と１の間で正規化された）で測定された時、発展させられた情報豊富なフイーチャーがないなら、最も適当な次の過程は、有用でローバストなモデルを発展させるよう努める代わりに該データへ戻ることである。

【０１６１】

時々、しかしながら、該第１発展段階のもう１つの成り行きがあり得る。該データから際立ったフイーチャーが発展することがあるかも知れない。このフイーチャーは極端に情報豊富で、事実、手元の問題用の”遺伝的コード（genetic code）”を表すかも知れない。この様な場合、より大きなデータ集合が該際立った遺伝子によりコード化された入力を使用して構文解析され得て（can be parsd）、この減少したデータ集合は、下にある法則を説明する数学的表現を発展させるために、遺伝的プログラミングフレームワーク内への入力として使用出来る。遺伝的プログラミングは、例えば、１９９４年発行、エムアイテープレス（M.I.T. Pres）、コザ、ジェイ．アール．（Koza, J.R.）著、”遺伝的プログラミング－自然的選択によるコンピユータのプログラミングについて（Genetic Programming-On the Programming of Computors by Natural Selection）”で説明されている。この表現は研究される過程の解析的説明を表し、発展型発見過程の最後の結果である。この過程を用いて、情報理論と発展の組み合わせは、見かけは混乱したシステム内の下にある秩序を閉じ込める数学的表現を発見することに帰着する。情報コンテントのためにフイーチャーを調べ、次いで実験型モデリングか、数学的発見か、又は該データに戻るか何れかに乗り込む、全体の過程はデータにドライブされるパラダイムに基づく”発見の科学（Science of Discovery）”への体系的取り組みを説明する。

【０１６２】

混乱したシステムの数学的説明の発展は基本的に内挿的性質（interpolative nature）か外挿的性質（extrapolative nature）へと該実験型モデルを変換する。かくして数学的表現は、該実験型モデルの開発で使用されるトレーニング集合の範囲の外側でデータ定義域

内に於いてさえ出力値を予測するため使用出来る。又数学的説明はモデル化されつつある過程又はシステム内への基本的見通しと恐らくは下にある原理の発見とを得るための励まし（stimulus）を提供する。

【０１６３】

【例】

均質ポリマー連鎖反応（POLYMER CHAIN REACTION ）｛ピーシーアール（PCR）｝フラグメントの同定

本発明が均質ピーシーアールフラグメントの同定に適用された。本方法は最初にデーエヌエイ溶解カーブ（DNA melting curve）の情報豊富な部分を同定し、次いで該入力スペクトラムの情報豊富な部分集合を使用して最適モデルを発展させる。

背景

デーエヌエイフラグメント同定は伝統的にゲル電気泳動（gel electrophoresis）により行われて来た。挿入染料（intercalated dyes）を使用する代替え方法はあり得る時間と感度での利点を提案している。この方法は、加熱時２重螺旋デーエヌエイが変性する（捲きほごれる）と該染料蛍光量（dye fluorescence）が減少することの観察に基づいている。温度に対する蛍光量をプロットする、最終のいわゆる”溶解曲線（melt curve）”のデータ解析は該デーエヌエイフラグメントのユニークな同定のベースを提供する。しかしながら、該方法は、特定的デーエヌエイフラグメントの精確な同定を、他の非特定的フラグメントの存在及び背景基盤（background matrix）からの蛍光ノイズの存在の両場合で、要求している。

スパイク（spiked）される食料サンプルの準備

この研究はピーシーアールを禁ずる知られる食料を評価した。該評価は、該禁止食料の禁止効果を克服するために、該反応へのウシ血清アルブミン（bovine serum alubumin）｛ビーエスエイ（BSA）｝の添加能力をテストした。加えて、溶解曲線解析を使用したピーシーアール製品の均質性検出が臭化エチジウム染色（ethidium bromide staining）を有する標準的ゲル電気泳動と比較された。

【０１６４】

食料は地域の食料雑貨店で購入され、４℃で貯蔵された。３０の異なる食料がビーエイエム（BAM）手順で事前強化（per-enriched）された。処方された強化法（enrichment）に従い、サンプルはサルモネラニューポート（Salmonella newport）でスパイクされるか又はスパイクされずに残されたが、表ＩＩＩ参照。該強化は次いでビーエイチアイ（BHI）｛デーアイエフシーオー（Difco）｝内で１：１０に薄められ、次いで３７℃で３時間培養された。

【０１６５】

【表２】

表 1

食料	事前強化ブロス (Broth)	食料：ブロス 希釈	接種 (Inoculation) レベル
アーモンド	エルビー (LB)	1：10	0, 10^4/mL, 10^5/mL
液状卵	テーエスビー (TSB)	1：10	0, 10^4/mL, 10^5/mL
赤小麦ぬか (Red Wheat Bran)	エルビー	1：10	0, 10^4/mL, 10^5/mL
ピーナッツバター	エルビー	1：10	0, 10^4/mL, 10^5/mL
ウオールナット	エルビー	1：10	0, 10^4/mL, 10^5/mL
挽きコーヒー	エルビー	1：10	0, 10^7/ml
インスタントコーヒー	エルビー	1：10	0, 10^7/ml
インスタント茶	エルビー	1：10	0, 10^7/ml
タイム (Thyme)	テーエスビー	1：10	10^7/ml
チョコレートアイスクリーム	脱脂粉乳 (Non fat dry milk)	1：10	10^7/ml
バジル	テーエスビー	1：10	10^7/ml
ホットチョコレートミックス	脱脂粉乳	1：10	10^7/ml
オレガノ (Oregano)	テーエスビー	1：100	10^7/ml
ペーストリナットミックス (Pastry Nut Mix)	エルビー	1：10	10^7/ml
全スパイス	テーエスビー	1：100	10^7/ml
ローズマリー	テーエスビー	1：10	10^7/ml
シナモン	テーエスビー	1：100	10^7/ml
小麦ぬか	エルビー	1：10	10^7/ml
カーネーション、ホットココアミックス	脱脂粉乳	1：10	0, 10^7/ml

【０１６６】
【表３】

食料	事前強化ブロス（Broth）	食料：ブロス　希釈	接種（Inoculation）レベル
ネスルのココア	脱脂粉乳	１：１０	0, 10^7/ml
オレオのクラム（Oreo Crumb）	脱脂粉乳	１：１０	0, 10^7/ml
スイスモカコーヒー	脱脂粉乳	１：１０	0, 10^7/ml
ネスルチョコレートリカー	脱脂粉乳	１：１０	0, 10^7/ml
ミルクチョコレート	脱脂粉乳	１：１０	0, 10^7/ml
ハーシーのココア	脱脂粉乳	１：１０	0, 10^7/ml
ダークココア（Dark Cocoa）	脱脂粉乳	１：１０	0, 10^7/ml
ウインナチョコレートコーヒー（Viennese Chocolate Cafe）	脱脂粉乳	１：１０	0, 10^7/ml
ウオールナットホイップ（Walnut Whip）	脱脂粉乳	１：１０	0, 10^7/ml
ネスルのミルクチョコレートクラム（Nestle's milk chocolate crums）	脱脂粉乳	１：１０	0, 10^7/ml

【０１６７】

ポリビニルポリピロリドン（Polyvinylpolypyrrolidone）｛ピーブイピーピー（PVPP）｝処理

グローバックサンプル（growback）の５００マイクロリットル（500 ul）のアリコート（aliquot）がピーブイピーピー｛クアリコン社（Qualicon, Inc.）｝の５０ｍｇのタブレットを含むチューブに追加された。該チューブはボルテックス（vortexed）されそして該ピーブイピーピーは１５分間澄むようにされた。最終浮遊物は次いで溶解過程で使用される。

サルモネラサンプルの準備

２ｍｌのスクリューカップチューブ（screw cup tube）で、強化すなわちピーブイピーピー処理サンプルの５マイクロリットルがデーエヌエイ挿入染料エスワイビーアールグリーン（DNA intercalating dye SYBR^R Green）｛モレキュラープローブ（Molecular Probes）｝の１：１０、０００希釈を含む溶解試薬｛５ｍｌビーエイエックス溶解バッファー（5ml BAX^R lysis buffer）と６２．５ｕｌ（マイクロリットル）ビーエイエックスプロテアーゼ（62.5 ul BAX^R Protease）｝の２００ｕｌ（マイクロリットル）に加えられた。該チューブは３７℃で２０分間次いで９５℃で１０分間培養された。９５℃の培養の後、４ｍｇ／ｍｌのビーエスエイ（BSA）溶液の５０ｕｌ（マイクロリットル）が該溶菌液（lysate）に追加された。これはピーブイピーピー処理済みと未処理のサンプルに行われた。対照として、幾つかのサンプル未処理で残された。この未精製バクテリヤ溶菌液の５０マイクロリットルが、パーキンエルマー７７００シークエンスデテクター計器（Perkin Elmer 7700 Sequence Detector instrument）で使用されるピーシーアールチューブ内に含

まれた１つのビーエイエックスサルモネラサンプルタブレット（BAX^R Salmonella sample tablet）を水和するため使用された。該チューブはキャップを付けられ、パーキンエルマー９６００サーマルサイクラー（Perkin Elmer 9600 thermal cycler）内で次のプロトコルに依り熱サイクルにかけられた。

【０１６８】

９４℃　　２．０分　　１サイクル
９４℃　　１５秒　　３５サイクル
７２℃　　３．０分
７２℃　　７分　　１サイクル
４℃　　”長期間（forever）”

増幅後分析（Post Amplification Analysis）

増幅後、下記条件で運転することによりパーキンエルマー７７００デーエヌエイシークエンスデテクター（Perkin Elmer 7700 DNA Sequence Detector）上で該溶解曲線が作られた。

【０１６９】

プレートの種類：　シングルリポーター（Single Reporter）
器械：　７７００シークエンスデテクションシステム（7700 Sequence Detection System）
運転：　実時間
染料層：　エフエイエム（FAM）
サンプルの種類：　未知である
サンプル容積：　５０ｕｌ（マイクロリットル）
運転条件：
７０℃　２分１サイクル　データ収集せず
６８℃　１０秒９８サイクル　データ収集する
自動インクレメント　＋０．３℃／サイクル
２５℃　”長期間”

該多成分データは該器械から移出され該分析に使用された。特定のデーエヌエイフラグメントの製作は該アンプリファイ（amplified）されたサンプルにビーエイエックスローディングダイ（BAX^R Loading Dye）の１５マイクロリットルを添加することにより検証された。次いで１５マイクロリットルのアリコートが臭化エチジウムを含む２％アガロースゲル（agarose gel）のウエル（well）内に装填された。該ゲルは３０分間１８０ボルトで運転された。特定の生成物は次いでユーブイトランスイルミネーション（UV transillumination）を使用して可視化された。

データ分析

生の蛍光量（raw fluorescence）データが処理用にマイクロソフトエクセル（Microsoft Excel）に移入された。この段階からデータを可視化し該データから予測をするため分岐的取り組みが使用された。

データ事前処理（Data Preprocessing）

蛍光ノイズを減らすために該データを事前処理することは成功するモデリングの尤度（likelihood）を増すことが実験的に決定された。該データ事前処理は次の過程から成り、すなわち、

ａ．蛍光データ（fluorescence data）の正規化、
ｂ．０．１℃の解像度でキュービックスプライン関数（cubic spline function）を用いた該正規化蛍光の内挿補間、
ｃ．内挿補間された蛍光スペクトラムの対数を取る、
ｄ．２５点サビツスキーゴレイ平滑化関数（25 point Savitsky Golay smoothing function）を用いた該蛍光の対数の平滑化、
である。

【０１７０】

最終温度スペクトラムはここで説明されるモデリング方法への入力の集合として使用される。該温度スペクトラムを使用した２つの異なるモデリング例を説明する。
過程ａ．データの正規化と可視化
該蛍光データは、最初にスペクトラム内の最低測定蛍光レベルを決定し、この値を、直流オフセットを除くために、該スペクトラム内の各点から引くことにより正規化される。上記の過程ａ．の正規化されたデータは次いでサビツスキーゴレイの平滑化アルゴリズム(Savitzky-Golay smoothing algorithm)で平滑化される。温度に対する平滑化蛍光の負の導関数｛－ｄｌｏｇ（Ｆ）／ｄＴ｝が取られ、－ｄｌｏｇ（Ｆ）／ｄＴ（ｙ軸）対温度（ｘ軸）としてプロットされる。
過程ｂ．該データからの予測
該正規化されたデータからスタートして、キュービックスプライン内挿関数（cubic spline interpolating function）を使用して０．１Ｃ分解能で該データは内挿補間される。次いで該内挿されたデータの対数が取られ、次いで２．５度（すなわち０．１℃で２５の点）上でサビツスキーゴレイの平滑化アルゴリズムを用いて平滑化される。温度に対する該ログの蛍光の負の導関数が取られ｛－ｄ（ｌｏｇＦ）／ｄＴ｝、サルモネラ用データ範囲：８２．０℃－９３．０℃（１２データ点）を用いて１．０Ｃ間隔でパース（parsed）された。
【０１７１】
方法比較用に、ここに説明された方法は２つの他の良く知られたモデリング方法：ニューラルネットワーク及びロジスティック回帰（logistic regression）、と比較され、結果は下表で報告される。
【０１７２】
見出された最も有効なＤＮＡフラグメント同定法は２つのモデリングスキームをシーケンシャルな仕方で背中合わせで使うことを含んでいる。同定の第１レベルはスメア（smear）を非スメア（non-smear）から分離することである。これに、非スメアサンプル用に関心のある特定のデーエヌエイフラグメントを同定することが続く。実際は、この階層的方法は、起こり得る出力カテゴリーを表す正、負そしてスメアを有する１つの３状態モデルを使用するより精確であった。
１．特定ピーシーアールフラグメントに対する非特定ピーシーアールフラグメントのモデリング
該ピーシーアールアンプリフイケーション過程（PCR amplification process）は、関心のあるデーエヌエイの特定の種類に対応するフラグメントのみならず非特定ピーシーアールフラグメントも作る。第１例は本方法の該非特定と特定のピーシーアールフラグメント間を区別する能力を展示する。１４９のロックされたプロセス（すなわち、対照）特定的トレーニングスペクトルと、問題食料（ピーシーアール用で問題があると知られる実際の食料）の３０９のテストスペクトルと、一緒に３０の非特定的又は”スメア”の蛍光スペクトルのグループが創られた。０．１℃の温度分解能を有して、１１１点を含む各サンプル用の温度スペクトル（１１．１℃の範囲上の）が創られた。該ロックされたプロセスと問題食料サンプルの両者が陽性と陰性の標本を含んだ。この例で、該陽性のサンプルは特定のバクテリヤ（例えば、サルモネラ）でスパイクされ（すなわち汚染され）そして陰性のサンプルはスパイクされぬ（汚染されぬ）ようにされた。該スメアサンプルはロックされたプロセストレーニング集合（１２スメアサンプル）と問題食料テスト集合（１８スメアサンプル）の両者にランダムに導入された。該陽性及び陰性の両サンプル状態は合併され２進のゼロ”０”文字でラベル付けされ、該スメアサンプル状態は２進の１”１”でラベル付けされた。
【０１７３】
ａ．入力の最も情報豊富な集合を発展させること
モデリング過程の第１歩は１１１次元の入力フイーチャー空間をより少ない、より情報豊富な部分集合に減じることである。前に説明した発展型フレームワークが該最も情報豊富なフイーチャーを発展させるために使用された。１００の遺伝子の初期遺伝子プールがラ

ンダムに発生され、そこでは各遺伝子は２進の１１１ビットの長さの記号列を有し、各ビットの状態は該対応入力フイーチャーが該遺伝子内で賦活されたかどうかを表している。該発展過程はセル当たり１サンプルとなるべき平均セル占有数（mean cell occupation number）により抑えられ、そして該発展は５世代より多く進んだ。各遺伝子の発展をドライブするために、グローバルエントロピー、又は適応度関数としてローカルエントロピーの数加重和（number-weighted-sum of local entropies）が使用された。該発展は固定サイズ化された部分範囲（すなわち、適応型ビニングよりむしろ、固定されたビン）を使用して進みそして該データは、上記説明の様に、０及び１の出力状態の数をバランスさせるようバランスさせられた。

【０１７４】

発展型過程を通して該１００の最も情報豊富な遺伝子のグローバルリストが保持された。全ての１１１の入力フイーチャーのビット頻度のヒストグラムが、発展した該情報豊富な遺伝子プール内で最も屡々発生するビットを同定するために、該発展の各世代の終わりで分析された。このヒストグラムはどの温度点が該出力状態に最も密接に付随したかについての情報を提供した。

【０１７５】

該１１１の点の温度範囲が０から１１０までインデックス（indexed）され、下記３１温度点が該発展型過程から選択された：１２，１４，１６，１８，２０，２２，２４，２６，２８，３０，３２，３４，３６，３８，４０，４２，４４，４６，５０，５２，５４，５６，５８，６０，６２，６４，８０，８２，８４，８６，８８。

【０１７６】

情報豊富な領域が該ヒストグラム内で観察されそしてこれらの領域に懸かる偶数番号インデックス点（上記リスト）が選択されたことは注意されるべきである。大抵の該選択された点が１２から６０の範囲に懸かることは注意されるべきである。これは該スメアサンプル用溶解曲線スペクトラムが該ベースライン上に立ち上がりそして該インデックス間隔［１２，６０］に対応する温度範囲内の陽性及び陰性両サンプルから別れ始めるからである。例えスメアがそれらの正に規定により可変溶解曲線構造を有するとは云え、主な構造的フイーチャーは該陽性のサンプル内よりも低い温度で一般に現れる。該陰性のサンプルは本質的に構造から自由である。かくして、本方法はより低い温度領域がスメアと非スメアの間の最良の区別が起こる場所であることを確認する。

【０１７７】

ｂ．パース（parsed）されたデータの全低次元射影のエグゾーストな探索

第１発展型過程で発見された該情報豊富な点を使って該トレーニングデータ集合がパースされた後、該減少したデータ集合は広いビニング範囲に亘り低次元でエグゾースチブに探索された。固定ビンとデータ集合バランシングが該エグゾースチブな過程を通して使用された。このモデリング問題で、次元当たり２６の固定ビンを使用して全２次元射影内への該３１次元入力空間の４６５の射影を発生することが該最良エグゾースチブモデルに帰着することが分かった。W_l^2＝１０、W_l＝５，定数項＝１のエントロピー加重係数が使用された。しかしながら、全４６５の射影を使用する該エグゾースチブモデルは、該射影の多くが情報より多くのノイズを導入するので、最適モデルであることを保証されない。それで、各ビットが該モデル用遺伝子プール内の与えられた２次元射影の包含（inclusion）（２進で１）と排除（exclusion）（２進で０）を表す４６５ビットの長さの２進記号列を使って第２の発展段階が行われた。

【０１７８】

ｃ．最良２次元モデルを発展させること

１００のランダム２進記号列が最初に発生されそしてそれらの適応度関数がテストデータ集合内誤差を該発展型過程をドライブする適応度関数として使用して計算された。該モデルは２０世代より多く発展させられそして最も情報豊富な遺伝子のグローバルなリストが保持された。最後に、この遺伝子プール内の最も情報豊富な遺伝子（最小テスト誤差に帰着する遺伝子に対応する）がスメア検出用遺伝子コードとして選択された。この遺伝子は

該包含２次元射影の１６３を有し残りの射影は排除された。これらの１６３の射影を使用した最小テスト誤差は該３２７テストケースから３つのエラー（3 errors out of the 327 test cases）（３０９問題食料サンプルと１８スメアサンプル）であって９９％より高いモデル精度に帰着する！

２．陰性のサンプルに対する特定のサルモネラピーシーアールフラグメント（陽性の）のモデリング

ピーシーアールモデリングの第２例として、本方法は食料サンプル内サルモネラに対応する特定のデーエヌエイフラグメントを同定するタスクを与えられた。もう１度、該ロックされた過程スペクトルが該トレーニングデータ集合として使用されそして該問題食料スペクトルが該テストデータ集合として使用された。上記説明のものと同様な過程が最良予測モデルを発展させるために使用された。

【０１７９】

ａ．入力の最も情報豊富な集合を発展させること

前の例で説明されたそれと同様な手順に従い、本方法は、下記の温度点：

１０，１３，１６，６１，６４，６７，７６，７９，８２，８５，８８，９１

に対応する１２入力フイーチャーの集合を発展させた。

【０１８０】

この例では、スペクトルの情報豊富な部分は該温度範囲のより高い端（点６１から９１の間）内にあることを注意する。これは余り驚くべきことではないが、それはポジテイブな（positive）溶解曲線内の主な構造が温度インデックス（temperature index）８０の周辺で起こるからである。

【０１８１】

ｂ．パースされたデータの全低次元射影のエグゾースチブな探索

第１発展過程で発見された該情報豊富な点を使用して該トレーニングデータ集合がパースされた後、減少したデータ集合は広いビニング範囲上で低次元でエグゾースチブに探索された。固定ビンとデータ集合バランシングが該エグゾースチブな過程を通して使用された。このモデリング問題で、次元当たり１９の固定ビンを使用した全３次元射影内への該１２次元入力空間の２２０の射影を発生することが最良エグゾースチブモデルに帰着することが分かった。前のサンプルでと同じエントロピー加重係数が使用された。この例で、全ての２２０の射影を使用することが最良モデルに帰着することが分かった。該２２０の射影の部分集合を発展させることは該テストデータ集合に関する予測精度を改良しなかった。全２２０の射影を用いて、該３０９の問題食料テストサンプル（スメアなしで）からの３０１が９７．４％の精度で適当と同定された。

結果

これらの実験中作られた該３０９のデータサンプルの中で、２０４はサルモネラでスパイクされそして１０５のサンプルが”ブランク（blank）”反応であった。該２０４のスパイクされたサンプルの中で、１４３のサンプルはアガロースゲルで陽性でありそして６１は該ゲルで陰性であった。該陰性のサンプルはピーシーアールの禁止か又は不適当なゲルか又はピーシーアール感度の結果と考えられ得る。該１０５の”ブランク”の反応の中で、９５は該ゲルに関し陰性で、そして１０は該ゲルに関し陽性であった。該陽性のサンプルは自然の食料汚染（例えば、液状卵サンプル）又は技術的誤りの結果と考えられ得る。

【０１８２】

下表は該３つのモデリング方法の結果を抄録する。該モデリング方法の各々の出力は１かゼロの間の数である。”１”はスパイクされた予測を表す一方”０”はスパイクされてない予測を表す。該数がゼロ又は１に近い程、該予測により高い信頼を置くことが出来る。０．５のしきい値より高いどんな予測も陽性と考えられた。下記方法の各々用数は期待予測と合致したサンプル数を示す

【０１８３】

【表４】

表 ＩＩ

	説明	期待される予測[2]	サンプル数[3]	本方法	ニューラルネット	ロジスチック回帰
スパイクされた／陽性ゲル	確認された陽性	１	１４３	１３９	１３８	１３４
スパイクされぬ／陰性ゲル	確認された陰性	０	９５	９３	９２	６４
スパイクされぬ／陽性ゲル	被汚染サンプル	１	１０	８	８	１０
スパイクされた／陰性ゲル[1]	検出感度	０／１	６１	５６／５	５５／６	４７／１４
		合計	３０９	３０１	２９９	２６９
			%合致度	９７．４１%	９６．７６%	８７．０６%

【０１８４】

[1]これらのサンプルはスパイクされたが、ゲル上では陰性であった。均質な検出はゲル検出より敏感なので、均質な検出で陽性のサンプルを検出するがゲルベースの方法では見出さないことが起こり得る。パーセント合致度計算時、このカテゴリーで全てのサンプルは正しいと仮定されている。

[2]”期待される予測”列はスパイクステイタスとゲル結果とに基づき１又は０を表示する。この数は該モデルが該トレーニングサンプルに基づき予測すると期待されたものである。

[3]”サンプル数”列は特定のスパイク／ゲルカテゴリーに分類されるサンプル数を表示する。

【０１８５】

本方法の階層化モデリングに加えて、ハイブリッドモデリングフレームワークが使われてもよい。

【０１８６】

ニューラルネットモデルは陽性／陰性の同定のみならずスメア／非スメアの同定用にも開発された。事実、より多くのデータが入手可能になると、多数のトレーニング／テストデータ集合が発生され得て多数ニューラルネット及びインフオエボルブテーエムモデル（InfoEvolve™ model）に帰着した。未知のサンプルは全てのモデルでテストされ得て個別モデル予測の統計に基づきカテゴリー化され得る。付録Ｇで論じる様に、この取り組みは、多数のデータ集合とモデリングパラダイムと上での多様化によりモデル偏倚のみならずデータ偏倚も減じる利点を有する。加えて、２つの別々のモデリング段階を続けて使用する

階層的取り組みはモデル精度を更に改善する。
ハイブリッドモデリング
本方法はデータモデリング用の強力なフレームワークを開示するが、どんなモデリングフレームワークも完全ではないことを注意することは大切である。全てのモデリング方法はその取り組みのためか又はデータに課されるジオメトリー（geometries）のためか何れかで、”モデル偏倚”を課す。本方法は追加的ジオメトリーの最小の使用を行いそして上記説明の様に幾つかの利点を有するが、しかしながら、本方法は基本的に外挿法的であるより寧ろ内挿法的である。比較的データの貧弱なシステムでは、この内挿法的特性は一般化の容易さを減じる。
【０１８７】
本方法の強さを利用しそしてその弱さを最小化するために、それはハイブリッドモデルを創るために他のモデリングパラダイムと組み合わされることが可能である。これらの他のパラダイムはニューラルネットワーク又は他の分類又はモデリングフレームワークであり得る。もし他のモデリングツール（含む複数ツール）が基本的に異なる哲学を有するなら、１つ以上の他のモデリングツール（含む複数ツール）を本方法と組み合わせることがモデル偏倚を平滑化する（smooth out）効果を有する。加えて、データ偏倚を平滑化するために異なるデータ集合を使用して各パラダイム内に多数のモデルが作られ得る。最後の予測結果は各モデルから来る個別予測の加重又は非加重の組み合わせとすることが出来る。ハイブリッドモデリングは多様なモデリング哲学の強さを利用するために極端に強力なフレームワークをモデリングに提供する。重要な意味で、この取り組みは実験型モデリングの究極の目標を表す。
【０１８８】
例えば、もし食料媒介病原菌用テスト（testing for foodborne pathogens）での上記説明例に於ける様に、偽陰性のパーセント（percento of false negative）を最小化したい望みがあるなら、該モデルのどれか１つがスパイクされたサンプルを予測したならば陽性の結果が報告されるであろう。もしこの規則がこの例のデータに適用されたなら、ゲル結果に基づく偽陽性（false positive）の率は０．７％より少なかったであろう。何れか１つのモデルについての偽陰性率はそれぞれ：本方法＝３．９％、ニューラルネットワーク＝４．５％そしてロジスチック回帰＝５．８％であった。
結論
この例は重要な実験型モデリング問題でのインフオエボルブテーエム（InfoEvolve^{TM}）のパワーを図解する。インフオエボルブテーエムは最初にデーエヌエイ溶解曲線の情報豊富な部分を同定し次いで該入力スペクトラムの情報豊富な部分集合を使用して最適モデルを発展させる。この例で追跡された一般的パラダイムは種々の産業及びビジネス応用品でテストされ大きな成功をもたらし、この新しい発見的フレームワークに強力な支持を提供している。
製造過程の例
ケルバーアール（Kelvar^{R}）製造過程での重要な変数は該ケルバーアールパルプ（Kelvar^{R} pulp）内に保持された残留湿気（residual moisture）である。該保持された湿気は該パルプの次の処理可能性と最終製品特性の両者に顕著な影響を有する。かくして最適制御戦略を規定するために該パルプ内の湿気保持に影響するキー要素、又はシステム入力を最初に同定することが重要である。製造システム過程は、乾燥処理用の全体の時間枠のために該入力変数と最終パルプ湿気間の多数の時間遅れの存在により複雑化される。パルプ乾燥処理のスプレッドシートモデルが創られ得るが、そこでは該入力は多くの前の時の幾つかの温度と機械的変数を表し、該出力変数は現在時刻のパルプ湿気である。最も情報豊富なフイーチャー組み合わせ（又は遺伝子）は、その変数の、より早期の時点でパルプ湿気に影響するのに最も情報豊富であるのはどの変数であるかを発見するためにここに説明された該インフオエボルブテーエム（InfoEvolveTM）を使用して発展させられ得る。
フロード（fraud）検出例
既知のフロード的（fraudulent）な場合のトレーニング集合を作るのが難しいからだけで

なく、フロードが多くの形式を取るかも知れないので、フロード検出は特に挑戦的応用である。フロードの検出は予測モデリングによりフロードを防止出来るビジネス用に可成りのコスト節約へ導き得る。フロードが起こる或るしきい値確率で決定出来る様なシステム入力の同定が望ましい。例えば、何が”ノーマル（normal）”な記録かを最初に決定することにより、或るしきい値より多く該ノーム（norm）から変化する記録が、より精密な精査用にフラグ建て（flagged）されてもよい。これは、クラスタリングアルゴリズムを適用し、次いでどのクラスターにも分類されない記録を調べることに依るか、又は各分野用の値の期待範囲を説明する規則を作ることに依るか、又は分野の異常な付随にフラグ建てすることにより行われてもよい。クレデイット会社は期待しない使用量パターン（usage patterns）にフラグを建てるこのフイーチャーをそれらの課金正式化過程内にルーチン的に組み込む。もしカード所有者（cardholder）が普通は彼／彼女のカードを航空券、レンタルカー、そしてレストラン用に使用するが、或る日それをステレオ機器か又は宝石を買うため使用するなら、その処理は、該カード所有者が彼のアイデンテイテイを検証する該カード発行会社の代表者と話を出来るまで、遅延してもよい。（参考文献：１９９７年発行、マイケル、ジェイ．エイ．ベリー、及びゴードン、リンホフ（Michael J. A. Berry, and Gordon Linhoff）著、”マーケッテイング、販売及び顧客サポート用データマイニング技術（Data Mining Techniques for Marketing, Sales, and customer Support）、７６ページ）。フロード検出でどの変数が最も情報豊富かを発見するために最も情報豊富なフイーチャー組み合わせ（又は遺伝子）がここで説明した本発明を使用して発展させられ得る。これらの変数は或る時間間隔に亘る購入の種類と量、クレデイットバランス、最近の住所変更他を含んでもよい。一旦入力の情報豊富な集合が同定されると、これらの入力を使用する実験型モデルは本発明を使用して発展させられ得る。これらのモデルは、フロード検出用の適合学習型フレームワークを創るために、新データが入ると規則的ベースで更新され得る。

マーケッテイング例

銀行は予防的アクションを行う時間を持つためにその要求払い預金勘定（demand deposit accounts）｛例えば、銀行当座預金（checking accounts）｝の顧客のアトリッション（attrition）の充分な警報を望む。それが余りに遅くなる前にトラブル範囲に見つけるために、起こり得る顧客のアトリッションをタイムリーな仕方で予測するキー要素又はシステム入力を決定することが重要である。かくして、勘定動向（account activity）の毎月の抄録はこの様なタイムリーな出力を提供しないが、処理レベルでの詳細データは提供するかも知れない。システム入力は、顧客が該銀行に置いて行く理由を含んでおり、この様な理由がもっともかどうかを決定するためにデータ源を同定し、次いで該データ源を処理経過データと組み合わせる。例えば、顧客の死亡が処理停止の出力を提供したり、或いは顧客は最早２週間毎に支払われないか又は最早直接預金を有せずかくして規則的な２週間ベースの直接預金は最早ない。しかしながら、内部決定で発生されたデータは処理データ内に反映されない。例は、該銀行がかって無料であったデビットカード処理用に今は課金しているから又は該顧客がローンのために拒絶されたから、顧客が去って行くことを含んでいる。｛１９９７年発行、マイケル、ジェイ．エイ．ベリー、及びゴードン、リンホフ（Michael J. A. Berry, and Gordon Linhoff）著、”マーケッテイング、販売及び顧客サポート用データマイニング技術（Data Mining Techniques for Marketing, Sales, and Customer Support）、８５ページ参照｝。予測的アトリッションを決定する中でどの変数が最も情報豊富であるかを発見するために、ここで説明した本発明を使用して最も情報豊富なフイーチャー組合わせ（又は遺伝子）が発展させられ得る。顧客属性のみならず銀行戦略に付随する内部管理も含めた両者が処理データパターンと組み合わされるデータベースを創ることは銀行戦略、顧客属性そして発見されるべき処理パターンの間の起こり得る情報豊富なリンケージを可能にする。これは今度は処理挙動を予測する顧客挙動予報モデル（customer behaviour forcasting model）の発展へ導くことが出来る。

金融予測例（Financial Forcasting Example）

金融予報｛例えば、株、オプション、ポートフオリオ（portfolio）そして物価指数（ind

ex pricing）｝での重要な考慮は株式市場の様な動的で移り気な活動場所では誤差の広いマージンを黙認する出力変数を決めることである。例えば、実際の物価レベルよりむしろダウジョンズ平均株価指数（Dow Jones Index）での変化を予測することは誤差のより広い許容限度（wider tolerance for error）を有する。一旦有用な出力変数が同定されると、次の過程は最適予測戦略を規定するために該選択された出力変数に影響するキー要素、又はシステム入力を同定することである。例えば、ダウジョンズ平均株価指数の変化はダウジョンズ平均株価指数での前の変化のみならず他に於ける国の及びグローバルの指数にも依存するかも知れない。加えて、グローバルな利率、外国為替レート及び他のマクロ経済的メザー（macroeconomic measures）が重要な役割を演ずる。加えて、最も金融的な予報問題は入力変数（例えば、前の価格変化）と終わりのタイムフレームでの最後の価格変化との間の多数の時間遅れの存在により複雑化する。かくして、該入力は前の多数の時刻での市場変数｛例えば、価格変化、市場の移り気（volatility of the market）、移り気モデルの変化（change in volatility model）、．．．｝を表しそして該出力変数は現在の時刻での該価格変化である。（参考文献：１９９６年発行、エドワードゲートレイ（Edward Gateley）著、”金融予測用ニューラルネットワーク（Neural Networks for Financial Forcasting）、２０ページ）。より早期の時期が指すどの変数が金融予測用市場変数への影響で最も情報豊富であるかを発見するためにここで説明する本発明を使用して最も情報豊富なフイーチャー組み合わせ（又は遺伝子）が発展させられ得る。一旦これら（変数、時点）の組み合わせが発見されると、それらは最適金融予測モデルを発展させるために使用出来る。

【０１８９】

下記はモデル発生にここで使用される説明した方法に関する擬コードリステイング（Pseude Code listing）である：

```
      LoadParameters();          //データ集合と、ビニングの種類の様な種
                                 々のパラメーターとをロードし、データ選出、
                                 エントロピー加重係数、データ部分集合の数
                                 他．．．をバランスさせる
Loop through subset_number {
  CreateDashSubset(filename)          //部分集合データをランダムに
  Loop through number of local models {
      EvolveFeatures();          //情報豊富な遺伝子を発展させる
      CreateTrainTestSubset();   //データ部分集合をトレーン／テスト部分
                                   集合に分ける
         EvolveModel();          //モデルを発展させる
      }
}
CreateDataSubset
      DetermineRangesofInputs;
      if(BalanceStatsPerCatFlag is TRUE)
         BalanceRandomize;
      else
         NaturalRandomize;
DetermineRangeofInputs
                  Loop through data records {
           Loop through input features {
                if(input feature value=max
                     or input feature value=min {
                  LoadMinMaxArray (feature index, feature value) ;
                  UpdateMinMax (feature value) ;
                }
           }             //入力フイーチャーループ終了
      }                   //データループ終了

BalanceRandomize
／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
／データ集合を現在の部分集合と残りの部分集合とに分ける；
／出力カテゴリー当たりの項目の数をユーザーが指定する。
／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
```

```
Loop through output stats {
    InitializeCountingState(output) to 0 ;
    InitializeCountingRemainingState(output) to 0 ;
}
Loop through data records {
    Set IncludeTrainFlag to FALSE ;
    Loop through input features {
        if(input features =min)  {
             if(input FeatureMinFlag=CLEAR){
                   IncludeTrainFlag=TRUE;
                   FeatureMaxFlag  =SET;
             }
        }
        elseif (input feature=max)  {
                  if (input FeatureMaxFlag=CLEAR)  {
                  IncludeTrainFlag=TRUE;
                  FeatureMaxFlag   =SET;
             }
         }
     }                         ／／フイーチャーループ終了
     output＝ReadOutputState;  ／／記録用に出力状態を読み出す
     guess＝GuessRandomvalue;
Threshold(output)＝NUMITEMSPERCAT／TotalCountinState (output)
                               ／／TotalCoutinState (output) は出力カテ
                                 ゴリー内の＃データ項目を意味する
```

／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
もしデータ記録がフイーチャー最小又は最大値の最初の場合なら、現在のデータ部分集合と残りのデータ部分集合の両者へ記録をコピーする。
／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

```
    if (IncludeTrainFlag=TRUE)  {    ／／現在の部分集合と残りのデータ部
                                         分集合の両者へ記録をコピー
         CopyRecordtoCurrentDataSubset;
         IncrementCountinState (output)  ;
         CopyRecordtoRemainingDataSubset;
         IncrementCountinRemainingState (output)  ;
    }
```

／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
或いは他にもし該出力カテゴリーの項目の数が過剰にNOTであるなら、該データ項目を該REMAININGデータ部分集合内に置き換える。
／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

```
elseif(Threshold (output) >MINIMUM__THRESHOLD) {
     CopyRecordtoRemainingData;
     IncrementCountinRemainingState (output) ;
     if (CountinState(output) <NUMITEMSPERCAT) {
          CopyRecordtoDataSubset;
          IncrementCountinState (output) ;
          }
  }
    ／／MINIMUM__THRESHOLDは、もう１つの現在の部分集合を創るために
     ／残りのデータ部分集合内に充分なデータが残ることを保証する
     ／よう典型的に０．５である
```

／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

或いは他にもし該ランダムな推定が該データ項目は現在のデータ部分集合へ行くべきと決めたなら、NUMITEMSPERCATの望まれる割り当てが越えられたかどうかをチェックして見る。もしそうでないなら、現在のデータ部分集合にデータ点を追加し、CountinStateをインクレメントする。

／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

```
elseif (guess<=  Threshold (output) ) {
     if (CountinState (output) <NUMITEMSPERCAT) {
          CopyRecordtoDataSubset;
          IncrementCountinState (output) ;
     else {
          CopyRecordtoRemainingData;
          IncrementCountinRemainingState (output) ;
          }
     }
```

／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

又は最後に、もし該ランダムな推定が該データ項目が該残りのデータ部分集合内に行くべきことを決めるならば、該残りの部分集合用割り当てが越えられたかどうかをチェックする。もしそうでないなら、該残りのデータ部分集合へ該データ項目を追加する。もし該割り当てが越えられたなら、もしそのカテゴリー内でより多くの項目が必要なら該データ項目を該現在のデータ部分集合に追加する。

／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

```
   elseif(CountinRemainingState (output) ＜ (1-Threshold (output) ) *
            TotalCountinState (output) )  {
            CopyRecordtoRemainingDataSubset;
            IncrementCountinRemainingData (output)  ;
        }
   elseif(CountinState (output) ＜NUMITEMSPERCAT){
            CopyRecordtoDataSubset;
            IncrementCountinDataSubset (output)  ;
        }
}                                          //データ記録ループの終了
//BalanceRandomizeの終了
NaturalRandomize
SampleSize=NumberOfDataRecords/NumberOfModels;
Threshold=1-SampleSize/NumberOfRemainingDataRecords;
Loop through output state {
      InitializeCountinState (output)  to 0;
      InitializeCountinRemainingState (output)  to 0;
}
Loop through data records {
      Loop through input features {
           if (input feature＝min)  {
                 if (input FeatureMinFlag=CLEAR){
                      IncludeTrainFlag=TRUE;
                      FeatureMinFlag  =SET;
                 }
           }
          elseif (input feature＝max)  {
                     if (input FeatureMaxFlag=CLEAR){
                     IncludeTrainFlag=TRUE;

                     FeatureMaxFlag  =SET;
                }
          }
    }                                 //フイーチャーループ終了
    outpur＝ReadOutputState;    //記録用に出力状態を読み出す
    guess＝GuessRandomValue;
```

／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
もしデータ記録がフイーチャーの最小又は最大値の最初の場合なら、該データ部分集合及び残りのデータ部分集合の両者に記録をコピーする。
／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

```
   if (IncludeTrainFlag=TRUE) {           //該データ部分集合と該残り
                                          //のデータ集合との両者に記
                                          //録をコピーする
         CopyRecordtoCurrentDataSubset;
         CopyRecordtoRemainingDataSubset;
    }
```

／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
又はもし該ランダムな推定が該データ項目が該残りのデータ部分集合内に行くべきことを決めるなら、そのカテゴリー用に該残りの部分集合の統計的限界が越えられたかどうかをチェックする。もし越えられないならば、該残りのデータ部分集合に該データ項目を追加する。もし該割り当てが越えられたなら、該データ部分集合に該データ項目を追加する。

```
／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
  elseif (guess<= Threshold) {
          if (CountinRemainingState (output) <
               Threshold*TotalCountinState (output) )
               CopyRecordtoRemainingDataSubject;
          else
               CopyRecordtoCurrentDataSubject;
      }
／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
又はもし該ランダムな推定が該データ項目が現在のデータ部分集合内に入るべきことを決めるなら、そのカテゴリー用に該現在の部分集合の統計的限界が越えられたかどうかをチェックする。もしそうでないなら、該現在のデータ部分集合に該データ項目を追加する。もし該割り当てが越えられたなら、該残りのデータ部分集合に該データ項目を追加する。
／＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊
      else{
      if (CountinState (output) <
          (1-Threshold)*TotalCountinState){
          CopyRecordtoCurrentDataSubject;
      else
             CopyRecordtoRemainingDataSubject;
    }
}                                    //データ記録ループ終了
／NaturalRandomizeの終了
EvolveFeatures
        SelectRandomStackofGenes(N);
        Loop Through each gene in Stack {
／＊＊＊＊＊＊＊遺伝子から部分空間を創る＊＊＊＊＊＊＊＊＊＊＊＊／
           ReadParameters();
           ReadSubspaceAxesfromGene();
           if (AdaptiveNumberofBinsFlag=SET)
                CalculateAdaptiveNumbins;
           else
                UseNumBinsinParameterList;
           if (AdaptiveBinPositionsFlag=SET)
                CalculateAdaptiveBinPositions;
           else
                CalculateFixedBinPositions;
／＊＊＊＊＊＊＊＊＊：遺伝子から部分空間を創ることの終了＊＊＊＊＊＊＊＊／
```

```
            ProjectTrainDataintoSubspace;
            CalculateGlobalEntropyforSubspace;
        }                                   ／／遺伝子ループの終了
        EvolveGenesUsingGlobalEntropy();    ／／遺伝的アルゴリズム
}
CreateTrainTestSubsets
    DetermineRangesofInputs;
    RandomizeTrainTestSubsets;
RandomizeTrainTestSubsets
 {
Threshold=ReadThresholdfromParameterList;
Loop through data records in Data Subset {
     Loop through input features {
          if (iput feature=min)  {
               if  (input FeatureMinFlag=CLEAR){
                    IncludeTrainFlag=TRUE;
                    FeatureMinFlag   =SET;
               }
          }
          else{
               if (input feature=max)  {
                   if (input FeatureMaxFlag=CLEAR){
                   IncludeTrainFlag=TRUE;
                   FeatureMaxFlag   =SET;
               }
          }
       }                                ／／フイーチャーループの終了
       output=ReadOutputState;          ／／記録用に出力状態を読み出す
       guess=GuessRandomValue;
```

```
        if (guess<= Threshold){
                    if (CountinTrainDataSubset (output) <
                         Threshold (output) *TotalCountinState
                         OR IncludeTrainFlag=TRUE)
                         CopyRecordtoTrainDataSubset;
                    else
                         CopyRecordtoTestDataSubset;
                  }
               else{
               if (CountinTestDataSubset (output) <
                    (1-Threshold)*TotalCountinState (output)
                    AND IncludeTrainFlag=FALSE){
                    CopyRecordtoTestDataSubset;
               else
                    CopyRecordtoTrainDataSubset;
      }
}                                        //データ記録ループの終了
//RandomizeTrainTestSubsetsの終了
ModelEvolution
 {
      GenerateRandomStackofModelGenes();    //モデル遺伝子が遺伝子のク
                                            //ラスターであるランダムモ
                                            //デル遺伝子を発生させる
         Loop through each model gene in stack {
         CalculateMGFF();                  //モデル遺伝子適応度関数
                                           //{エムジーエフエフ(MGFF)}
                                           //の計算
           }                               //モデル遺伝子ループの終了
         EvolveFittestModelGene();         //最適モデル遺伝子を発展さ
```

```
                                        ／／せるため遺伝的アルゴリズム
                                        ／／をドライブするようエムジー
                                        ／／エフエフを使用
}
CalculateMGFF－モデル遺伝子適応度関数（エムジーエフエフ）の計算
 {
     IdentifyFeatureGenes();            ／／フイーチャー遺伝子の集合を
                                        ／／同定するためモデル遺伝子を
                                        ／／パース（parse）する
     Loop through each feature gene {
        CreateFeatureSubspace();
        Loop through each test record {
             ProjectTestRecordintoSubspace();
             UpdateTestRecordPrediction();
         }
     }
     Total__Error=0;
     Loop through each test record {
        if (RecordPrediction!=ActualRecordOutput)
           TotalError=TotalError+1;      ／／インクレメント誤差
     }
     MGFF=Total__Error;
}
```

本発明の好ましい実施例がここで説明された。付属する請求項により規定された本発明の真の範囲から離れることなく変更や変型が該実施例内で行われ得ることは勿論理解されるべきである。本実施例は好ましくは、コンピュータで実行可能なソフトウエア命令のセットとしてソウトウエアモジュール内で説明された方法を実施するロジックを含むのがよい。中央処理ユニット（”シーピーユー（CPU）”）、又はマイクロプロセサーは該トランシーバーの動作を制御する該ロジックを実行する。該マイクロプロセサーは説明された機能を提供するために当業者によりプログラムされ得るソフトウエアを実行する。

【０１９０】

該ソフトウエアは、磁気デイスク、光デイスク、そして該シーピーユーにより可読な何等かの他の揮発性［例えば、ランダムアクセスメモリー｛”ラム（RAM）”｝］又は不揮発性［例えば、読み出し専用メモリー｛”ロム（ROM）”｝］フアームウエア記憶システムを含むコンピュータ可読の媒体上に保持される２進のビットのシーケンスとして表され得る。データビットが保持される該メモリー配置も又該記憶されるデータビットに対応する特定の電気的、磁気的、光学式又は有機的特性を有する物理的配置を有している。ソフトウエア命令はメモリーシステムを有する該シーピーユーによりデータビットとして実行され、該電気信号表現の変換と該メモリーシステム内のメモリー位置でのデータビットの保持をもたらし、それにより該ユニットの動作を再構成させるか又は他の仕方に変えさせる。該実行可能なソフトウエアコードは、例えば、上記説明の様な方法を実施してもよい。

【０１９１】

ここで説明されたプログラム、過程、方法そして装置は、他のように指示されてない限り、どんな特定の種類のコンピュータ又はネットワーク装置（ハードウエア又はソフトウエア）にも関係付けられず、限定されないことは理解されるべきである。種々の種類の汎用又は専用コンピュータ装置又は計算装置がここで説明された開示に依って使用されてもよく、動作を行ってもよい。

【０１９２】

本発明の原理が適用される広範な種類の実施例を見ると、図解された実施例は単に例示的で本発明の範囲を限定すると取られるべきでないことを理解すべきである。例えば、本発

明は金融サービス市場、宣伝及びマーケッテイングサービス、製造過程に関連するシステム又は大きなデータ集合を有する他のシステムで使用されてもよい。加えて、該流れ線図の過程は説明されたものとは他のシーケンスで用いられてもよく、そして該ブロック線図ではより多く又はより少ない要素が使われてもよい。

【０１９３】

ハードウエア実施例は種々の異なる形式を取ってもよいことは理解されるべきである。該ハードウエアはカスタムゲートアレー（custom gate array）または特定用途向け集積回路（application specific integrated circuit）｛”エイシック（ASIC）”｝で集積回路として実施されてもよい。勿論、該実施例は個別ハードウエア部品（discrete hardware components）と回路で実施されてもよい。特に、ここに説明した論理構造と方法の過程はエイシックの様な専用ハードウエアで、又はマイクロプロセサー又は他の計算素子により行われるプログアム命令として実施されてもよい。

【０１９４】

請求項はその効果に対し述べられていない限り要素の説明された順序に限定されるとして読まれるべきでない。加えて、何れの請求項でも用語”手段（means）”の使用は３５ユー．エス．シー．§１１２、パラグラフ６を行使するよう意図されており、該用語”手段”を有しない何れの請求項もそのように意図されてない。従って、下記請求項の範囲と精神に入る全ての実施例とその等価物は本発明として請求されている。

【図面の簡単な説明】

【図１】　本方法の全体的流れを図解するブロック図である。

【図２Ａ及び２Ｂ】　適合型ビニングの例を示す。

【図２Ｃ】　データバランシングの方法を示す。

【図３Ａ】　１次元のフイーチャー部分空間を示す。

【図３Ｂ】　２次元のフイーチャー部分空間を示す。

【図３Ｃ】　３次元のフイーチャー部分空間を示す。

【図４】　どの入力がフイーチャー部分空間に含まれるかを表す例示的２進ビット記号列を示す。

【図５Ａ及び５Ｂ】　”情報豊富な”入力フイーチャーの発展を図解するブロック線図である。

【図５Ｃ】　２進記号列適応度の加重ルーレット選択ホイール（weighted roulette wheel）を示す。

【図５Ｄ】　交叉（crossover）操作線図を示す。

【図６】　ローカルエントロピーパラメーターを計算する方法を図解するブロック線図である。

【図７】　グローバルエントロピーパラメーターを計算する方法を図解するブロック線図である。

【図８】　ローカル及びグローバル情報コンテントの計算を図解する。

【図９】　ローカルエントロピーパラメーターとグローバルエントロピーパラメーターの例を示す。

【図１０Ａ】　最適モデルを決定する方法を図解するブロック線図である。

【図１０Ｂ】　モデル発展の方法を図解するブロック線図である。

【図１１】　情報写像（information map）を発生させる方法を図解する。

【図１２】　遺伝子リストとそれの付随情報写像の例である。

【図１３】　エグゾースチブな次元のモデリング過程の方法を図解するブロック線図である。

【図１４】　出力状態確率ベクトル／出力状態値を計算する過程の方法を図解するブロック線図である。

【図１５】　モデル遺伝子用適応度関数を計算する方法を図解するブロック線図である。

【図１６】　１つのフレームワークを発展させるために分布状階層的モデリングの方法を図解するブロック線図である。

【図１７Ａ及び１７Ｂ】　フレームワーク発展の方法を図解するブロック線図を含む。
【図１８Ａ】　スーパーフレームワークを発展させるための分布状モデリングの方法を図解するブロック線図である。
【図１８Ｂ】　スーパーフレームワーク発展用の考慮点のリストである。
【図１９Ａ及び１９Ｂ】　クラスター発展の方法を図解するブロック線図である。
【図１９Ｃ】　データクラスターを発見する方法を図解するブロック線図である。
【図１９Ｄ】　画像的表現用グローバルクラスタリング指数の計算方法を図解するブロック線図である。

【図１】

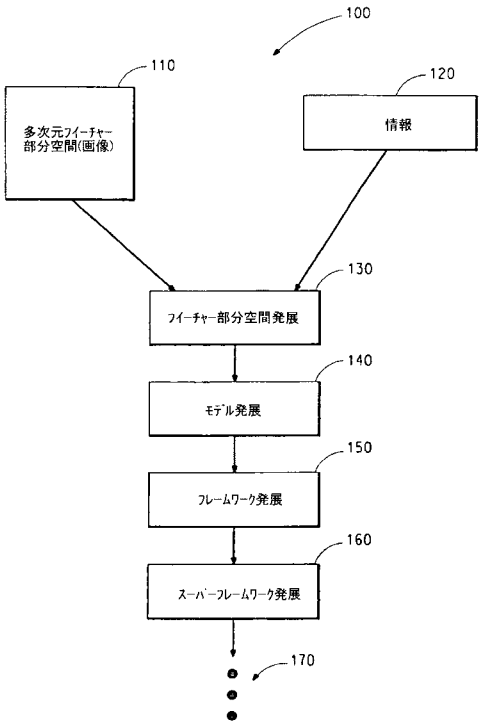
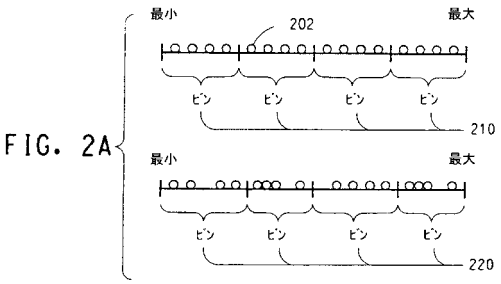


FIG. 1

【図２Ａ－Ｂ】



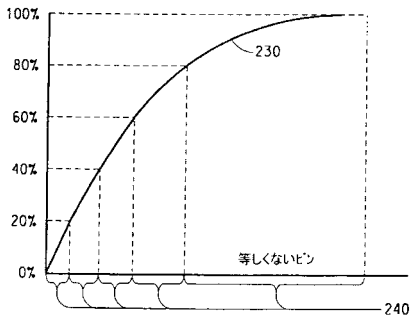


FIG. 2B

【図２Ｃ】

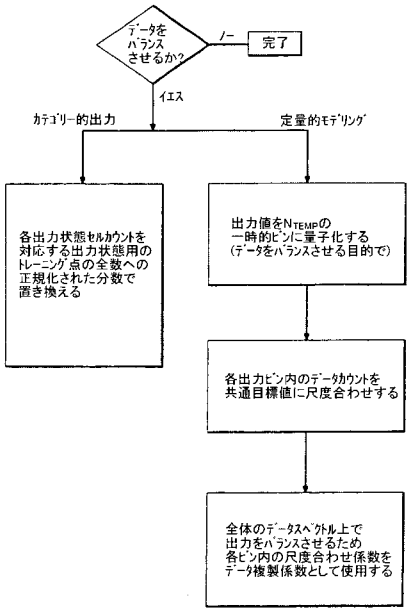


FIG. 2C

【図３Ａ－Ｂ】

FIG. 3A

1次元フィーチャー部分空間

データ	0010	0100
DP1 (.5, 1.5, .5, 1.5)	C1	C2
DP2 (.9, .7, 1.2, 1.2)	C2	C1
DP3 (.2, 1.9, 1.7, 1.1)	C2	C2
DP4 (.1, .4, .3, 1.7)	C1	C1

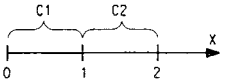
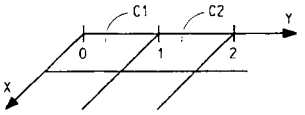


FIG. 3B

2次元フィーチャー部分空間

データ	1100	1010
DP1 (.5, 1.5, .5, 1.5)	C2	C1
DP2 (.9, .7, 1.2, 1.2)	C1	C2
DP3 (.2, 1.9, 1.7, 1.1)	C2	C2
DP4 (.1, .4, .3, 1.7)	C1	C1

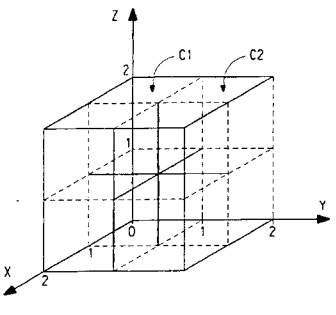


【図３Ｃ】

FIG. 3C

3次元フィーチャー部分空間

データ	1011	1101
DP1 (.5, 1.5, .5, 1.5)	C1	C2
DP2 (.9, .7, 1.2, 1.2)	C2	C1
DP3 (.2, 1.9, 1.7, 1.1)	C2	C2
DP4 (.1, .4, .3, 1.7)	C1	C1



【図４】

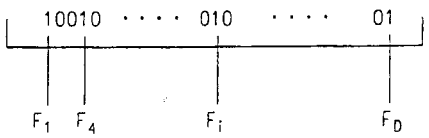


FIG. 4

【図５Ａ】

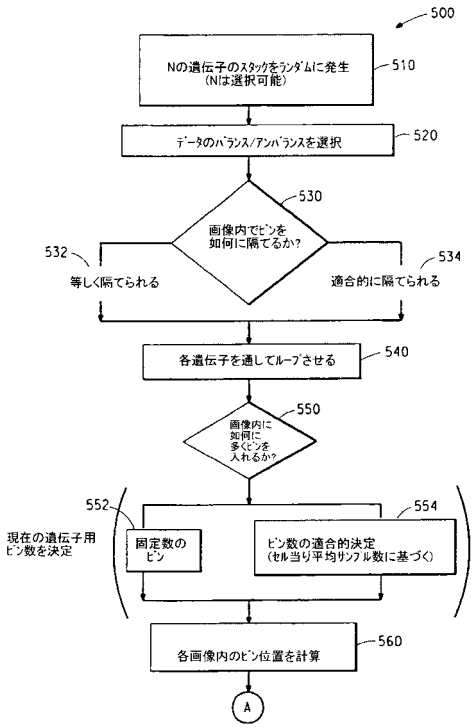


FIG. 5A

【図５Ｂ－Ｄ】

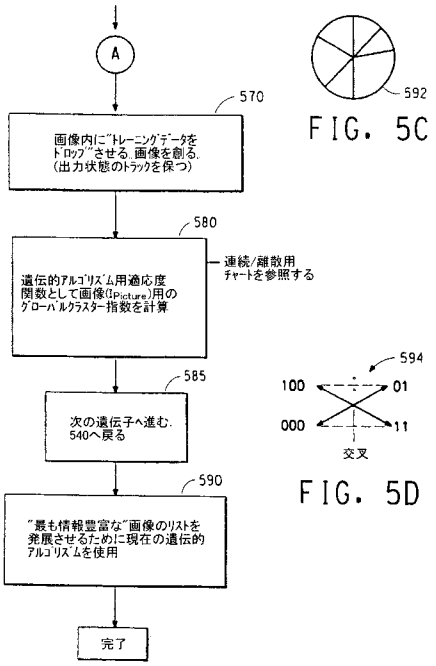


FIG. 5C

FIG. 5D

FIG. 5B

【図６】

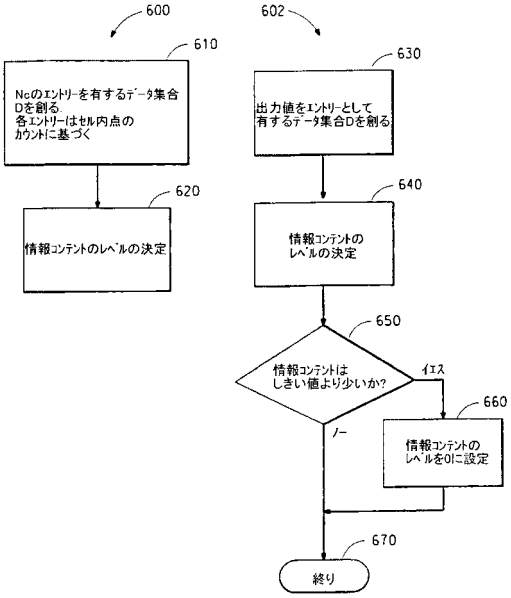


FIG. 6

【図７】

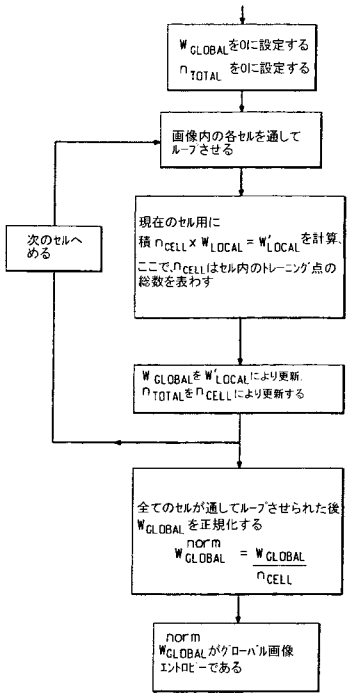


FIG. 7

【図８】

A 情報の定量評価の規定:

$D = \{d_1, \ldots\ldots, d_N\}$

i. もし該データ集合, Dが負のdiを有するなら下記過程に従う,
もしそうでないなら ii へスキップする

1. D(データ集合)のdmin番号を見出す
2. Dのすべての要素からdminを引く

$d_i = d_i - d_{min}$.

ii. $d_{total} = \sum_{i=1}^{N} d_i$

$F = \{ \frac{d_i}{d_{total}}, \ldots\ldots, \frac{d_N}{d_{total}} \}$, $f_i = \frac{d_i}{d_{total}}$

シャノン(1947-)はエントロピー関数を定義し, それはニシにより一般化された

** $S = \frac{\sum_{i=1}^{N} f_i \log f_i}{\log (\frac{1}{N})}$ (エントロピー関数)

ニシ ⟶

0と1の間の値を持つよう
正規化された

均一データの例:

全ての $d_i = d_o$ と仮定　　$D = \{d_o, \ldots\ldots, d_o\}$

$d_{total} = N d_o$　　均一データ

$f_i = \frac{d_o}{N d_o} = \frac{1}{N}$, $F = \{\frac{1}{N}, \ldots\ldots, \frac{1}{N}\}$

$S = (\text{エントロピー}) = \frac{\sum_{i=1}^{N} \frac{1}{N} \log (\frac{1}{N})}{\log (\frac{1}{N})} = 1!$

非常に不均一なデータの例

唯一つ d_i が正 > 0 これを d_o と呼ぶ

他は全て d_i's $= 0$

$D = \{d_o, 0, 0, \ldots\ldots\}$

$F = \{1, 0, \ldots\ldots\}$

$S = \frac{\log 1}{\log (\frac{1}{N})} = 0$

FIG. 8

【図９】

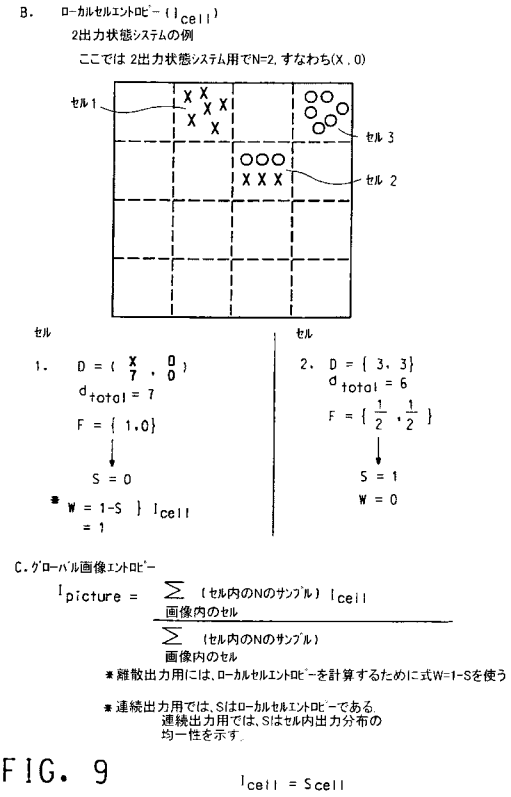


FIG. 9

【図１０Ａ】

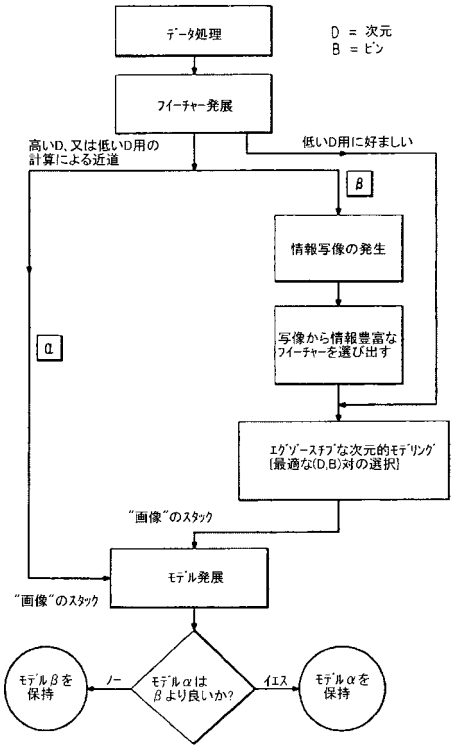


FIG. 10A

【図１０Ｂ】

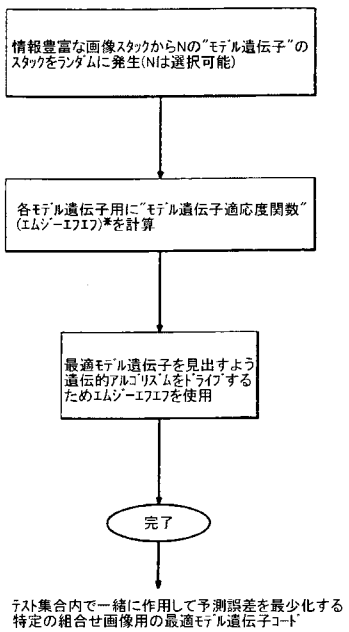


FIG. 10B

【図１１】

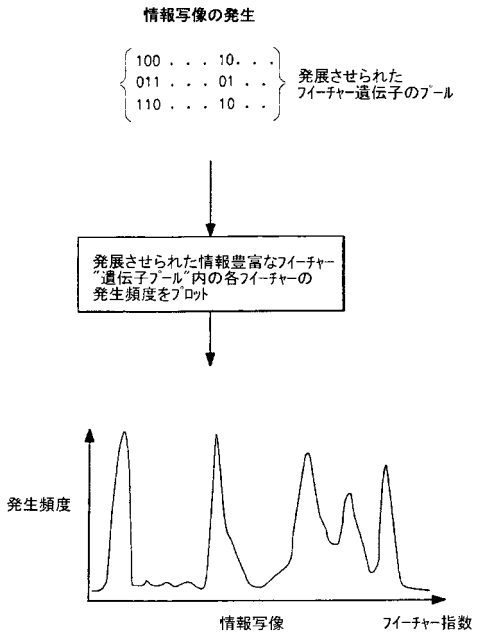
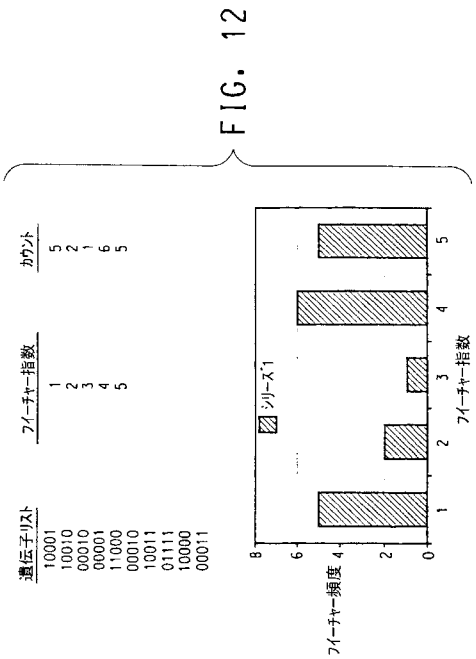


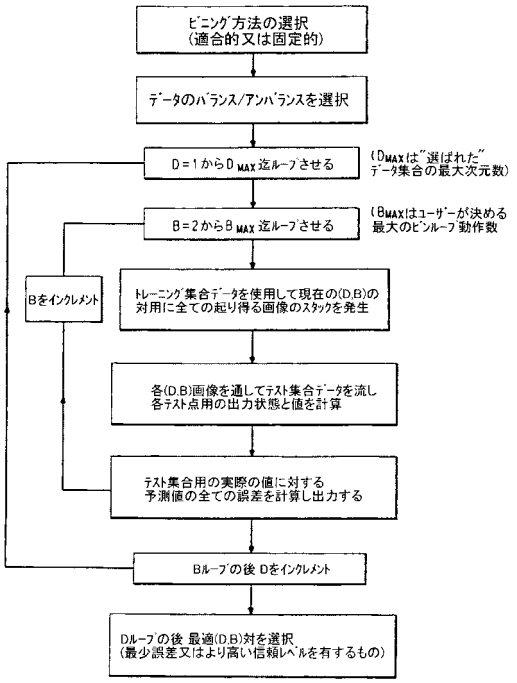
FIG. 11

【図１２】

FIG. 12



【図１３】



【図１４】

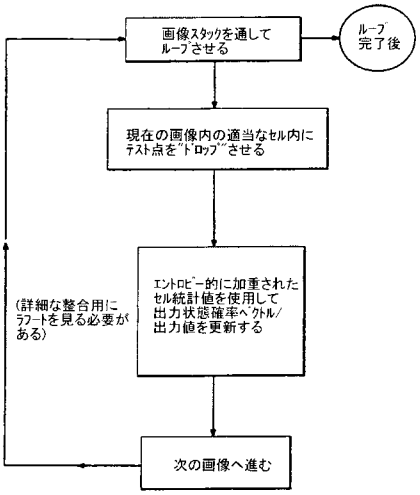


FIG. 14

【図１５】

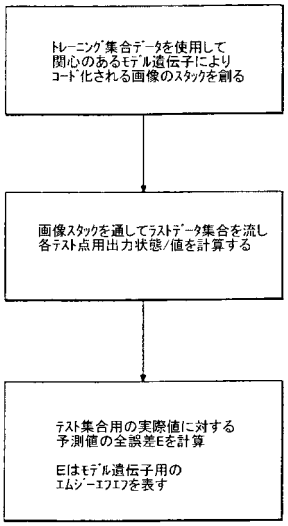


FIG. 15

【図１６】

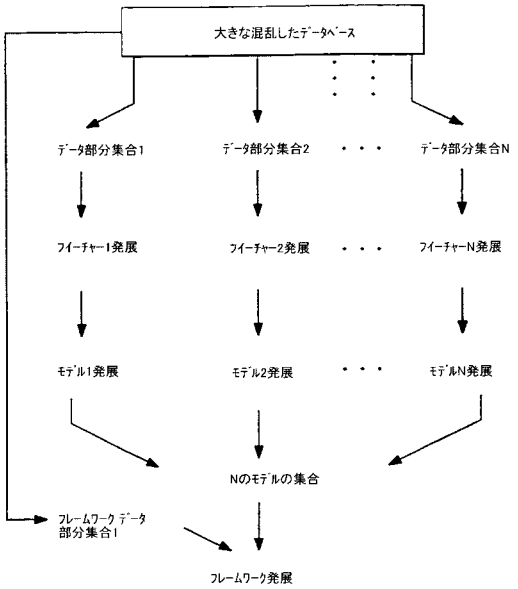


FIG. 16

【図１７Ａ】

フレームワーク発展法一の概要

1. Nのモデルの集合でスタートする。
2. Mのデータ項目を有するフレームワークデータ部分集合を使用して予測マトリックスを創る。
   予測マトリックスはM行と(N+1)列を有する。

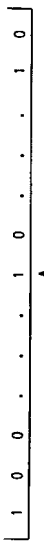
3. 予測マトリックスの各マトリックス要素は該フレームワークデータ部分集合用の予測値（カテゴリー的モデリング用には離散的な状態予測値、定量的には連続値）を有する。
4. 予測マトリックスの最後の列(第(N+1)の列)はフレームワークデータ部分集合用の実際の出力値を含む。
5. 該予測マトリックスはモデル組み合わせを発展させるため使用される データ表として使用される。

FIG. 17A

【図１７Ｂ】

フレームワーク発展(続き)

フレームワーク遺伝子はNビットを有する。



各ビットは1つのモデルを指す
それで該モデルはフレームワーク発展用の入力フィーチャーである。

＊＊＊＊＊

6. 次いでフレームワーク発展は、該データ表として該予測を、入力として該個別モデルを使用し、モデル発展用での図1に於ける様に進む。

重要な注意

フレームワーク発展では、モデルを発展させるためにトレーニング/テストのデータで生データに直接アクセスすることはない、該予測マトリックスはフレームワークを一般化するためフレームワークデータ部分集合を使用する。

フレームワーク＝モデル組み合わせの集合

FIG. 17B

【図１８Ａ】

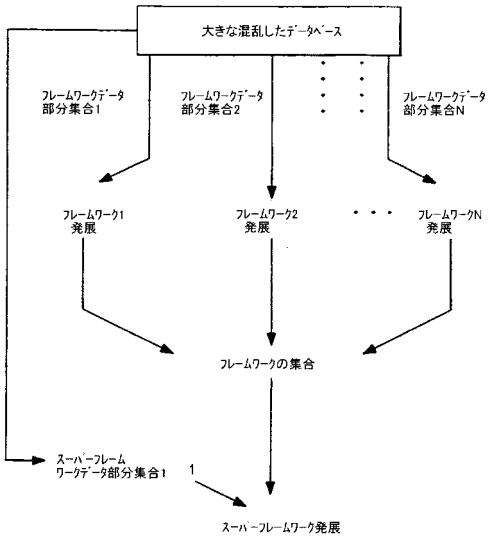


FIG. 18A

【図１８Ｂ】

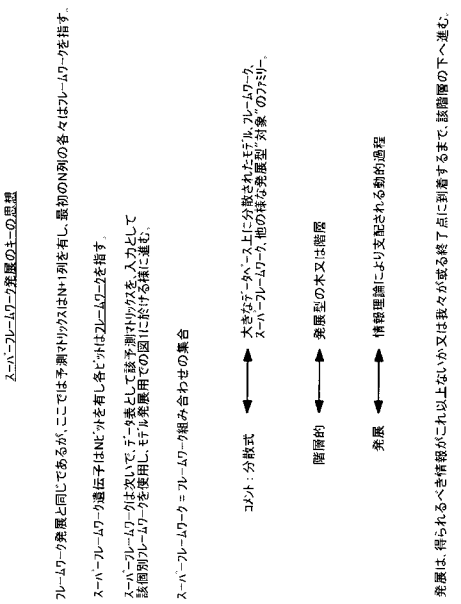


FIG. 18B

【図１９Ａ】

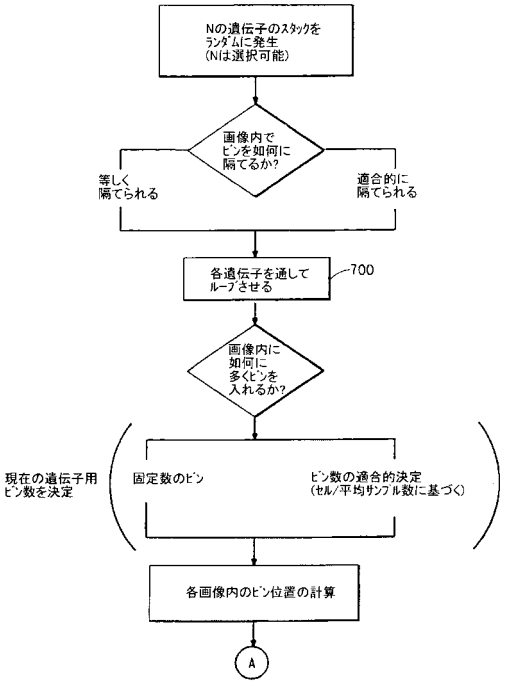


FIG. 19A

【図１９Ｂ】

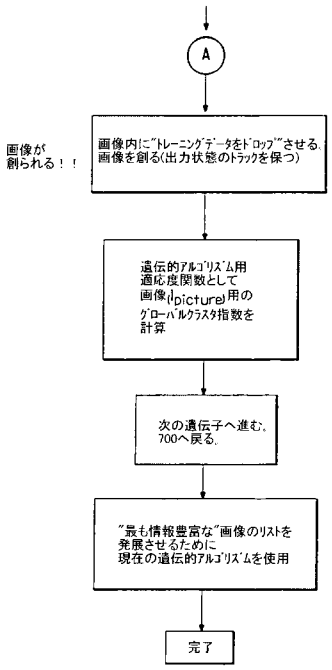


FIG. 19B

【図１９Ｃ】

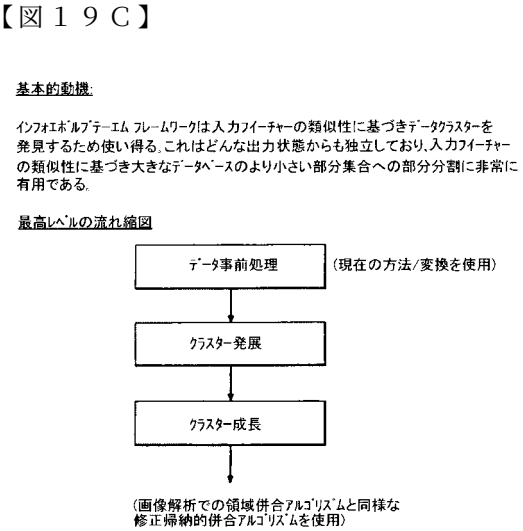


FIG. 19C

【図１９Ｄ】

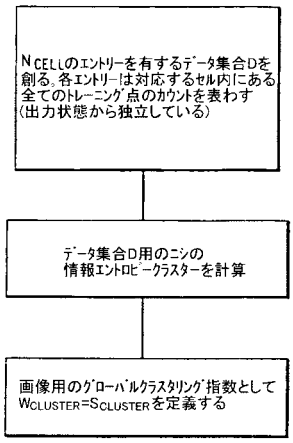

NCELLのエントリーを有するデータ集合Dを創る。各エントリーは対応するセル内にある全てのトレーニング点のカウントを表わす(出力状態から独立している)
データ集合D用のニシの情報エントロピークラスターを計算
画像用のグローバルクラスタリング指数としてWCLUSTER=SCLUSTERを定義する

FIG. 19D

フロントページの続き

(72)発明者　バイドヤナサン，アクヒルスウオー・ガネシユ
アメリカ合衆国デラウエア州１９７０７ホツケシン・ロビンコート４４
(72)発明者　オーエンス，アーロン・ジエイ
アメリカ合衆国デラウエア州１９７１３ニユーアーク・ルネイプレイン２３・シルバーブルツク
(72)発明者　ウイトコム，ジエイムズ・アーサー
アメリカ合衆国ノースカロライナ州２８７１２ブルバード・カントリークラブロード１３１５

審査官　長谷川　篤男

(56)参考文献　特開平１０－０９０００１（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06N　3/00
IEEE Xplore
JSTPlus(JDreamII)