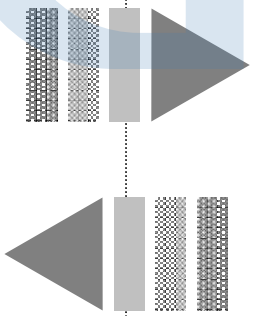
データマイニングと統計

マイニング ≠ 統計であって
マイニング　　統計ではない

伝統的な統計的解析手法

- 同一データからは，同一分析結果 （分析結果は一意に決まる）
- 多変量解析の手法であっても，原則的同一の分析結果
- オプションの違いによる多少の差異は あるものの，本質的な解釈まで変化することはない



データマイニング

- 同一データからは，同一分析結果が導かれるとは限らない （分析結果の不定性）
- 選択した手法の性質によって変化する

MediBic
株式会社メディビック

データマイニングのスタンス

「最適性，不定性」は短所なのだろうか？

- 「真の最適解を探索」⇔コストと時間がかかる
  - 変数の組み合わせが増加
  - 計算時間の指数関数的に増加
- 「目的を達するパフォーマンスの組合わせ」を探索
  - 知見の価値が手間（コスト）を上回れば成功

知見の価値と手間のバランスが重要

MediBic
株式会社メディビック

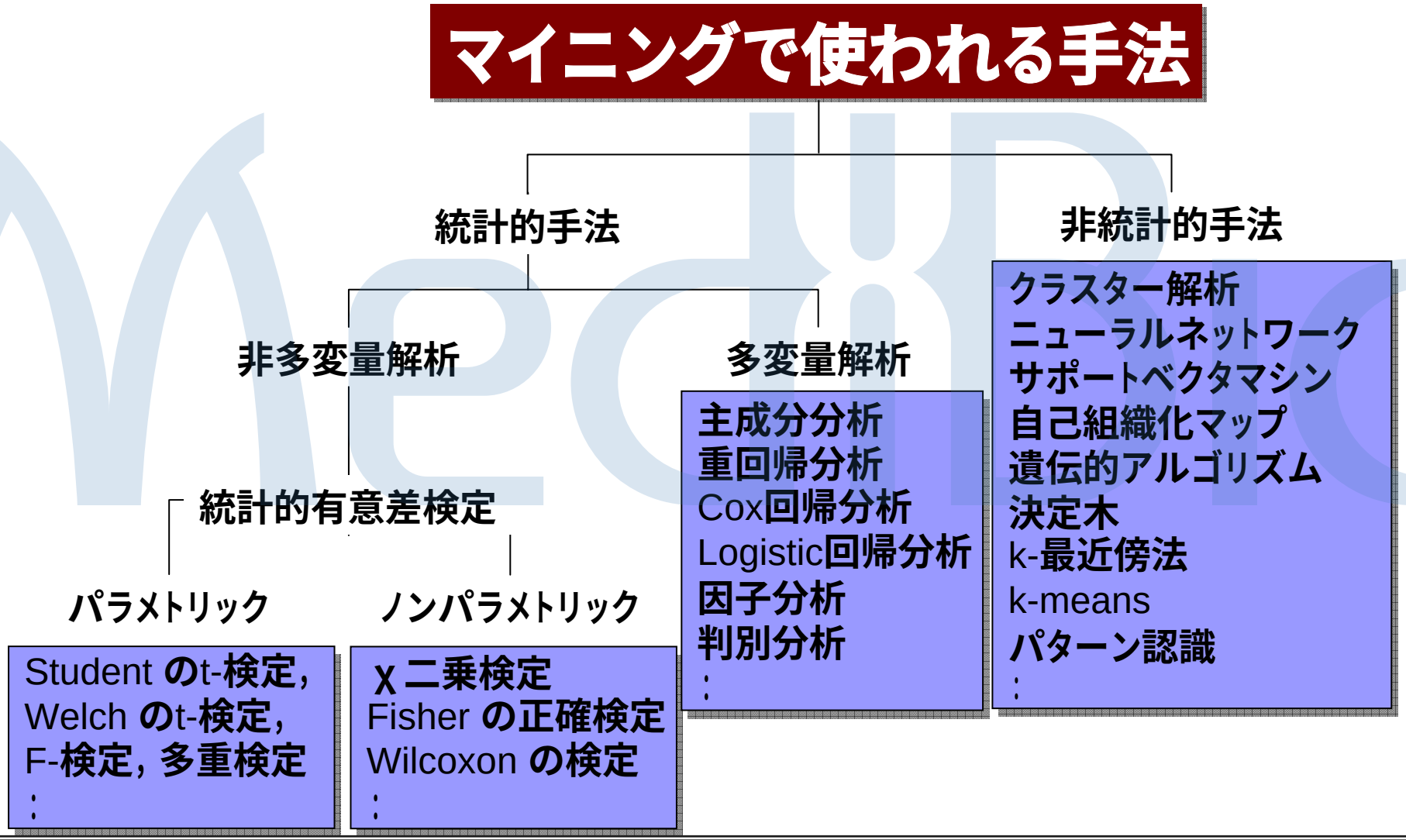
パターン認識と機械学習

■ パターン認識
観測した『認識対象がいくつかの概念に分類できる場合，観測されたパターンをそれらの概念のうちの１つに対応させる処理のことをいう』

■ 機械学習
コンピュータに学習機能を持たせること

分類	方法概要	手法
教師付き機械学習 Supervised learning	・予め分類やパターンの情報を与える ・学習成果をもとに，新たなデータを分類させる	・ニューラルネットワーク ・サポートベクタマシン ・ k-最近傍法 ・ 判別分析
教師なし機械学習 Unsupervised learning	・事前情報なし ・与えられたデータの類似性をもとに分類	・主成分分析 ・因子分析 ・クラスター解析 ・k-means ・自己組織化マップ

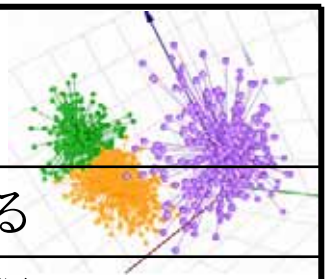
データマイニングで使われる手法例

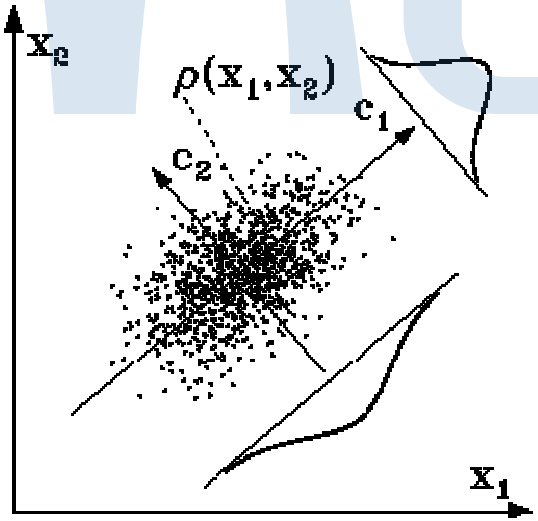


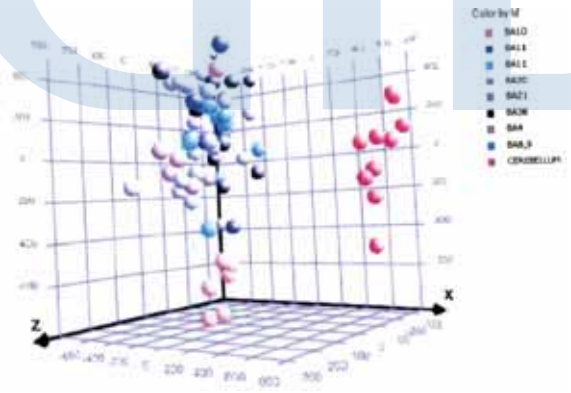
MediBic
株式会社メディビック

主成分分析 – Principal Component Analysis

目的	■データの傾向をつかむ ■データの分類基準をつかむ
長所	多次元の説明変数を，少数次元に落とし込むことができる
短所	結果が，常に解釈可能(既知の知見と整合性がある)とは限らない ▪ 既知の知見が不十分なために分析結果が正当に評価できない場合 ▪ 分析に使用した変数のセットが不適切な場合 （ 実態を把握するために必要な変数が抜け落ちているなど ）







“Molecular characterization of suicide by microarray analysis”
American Journal of Medical Genetics Part C: Seminars in Medical Genetics. Volume 133C, Issue 1，2005. Pages 48-56

MediBic 株式会社メディビック

クラスター解析– Cluster Analysis

目的	互いに似たものを集めてクラスタ(グループ)を作り，対象を分類すること	
種類	階層的クラスタリング Hierarchical clustering	・似たもの同士を併合して幾つかのグループにまとめる
	非階層的クラスタリング Non-hierarchical clustering	・似たものが結果的に同じグループに入るように集合を分割する

距離関数例：

- Euclidean
- Squared Euclidean
- Differential
- Pearson Absolute
- Angular （cos θ）

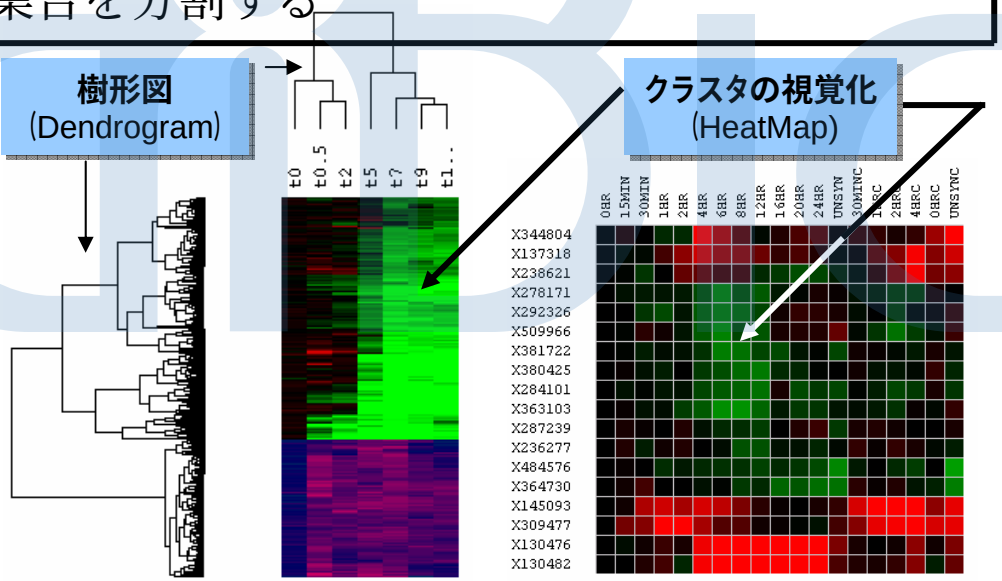
$$\sqrt{\sum_i (x_i - y_i)^2}$$

$$\sum_i (x_i - y_i)^2$$

$$\sqrt{\sum_i [(x_{i+1} - x_i) - (y_{i+1} - y_i)]^2}$$

$$\left| \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)(\sum_i (y_i - \bar{y})^2)}} \right|$$

$$\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}$$



クラスタ間距離手法例：

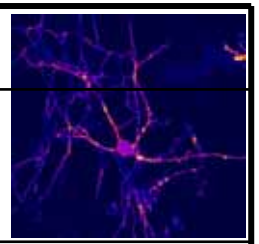
- Complete Linkage (furthest neighbor, **最長距離法**)
- Single Linkage (nearest neighbor, **最短距離法**)
- Average Linkage (group average, **群平均法**)
- Ward's Method （**ウォード法**）

分類感度高い．空間の拡散が起こりやすい．
分類感度低い．鎖状のクラスターを作る傾向がある．
分類感度中位．最長距離と最短距離の中間的．
分類感度高い．明確なクラスタを作る．

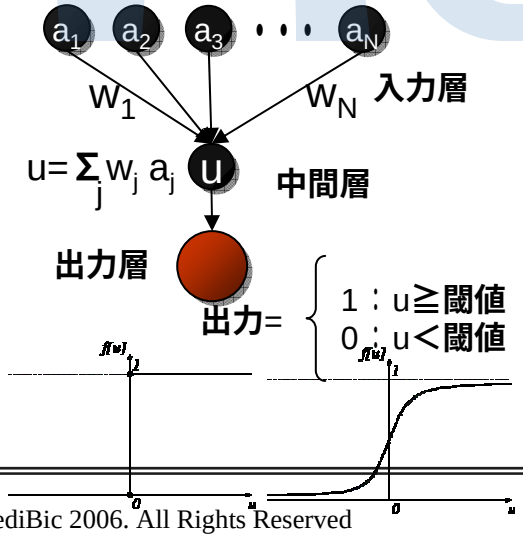
MediBic 株式会社メディビック

ニューラルネットワーク– Neural Network

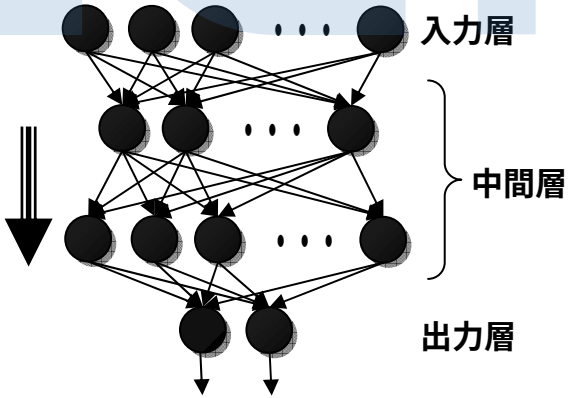
目的	分類や識別，予測，パターン認識	
概要	人間の脳の神経細胞をモデル化した手法．機械学習の一種 ・それぞれの接合強度を調節することにより「学習」する	
種類	階層型NN	単純パーセプトロンがいくつも層となり複雑な反応を示すようになったもの
	非階層型NN	① 相互結合型ネットワーク ② 競合学習型ベクトル量子化ニューラルネットワーク



単純パーセプトロン

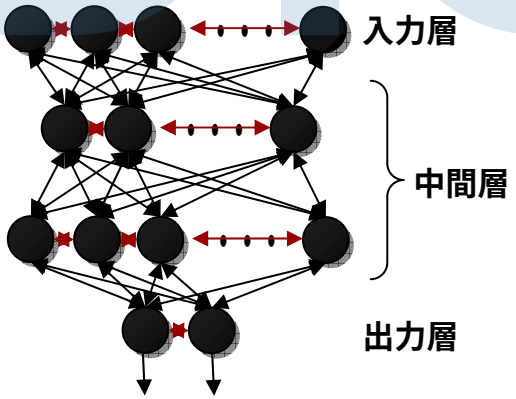


階層型ニューラルネットワーク



バックプロパゲーション法

相互結合型ネットワーク



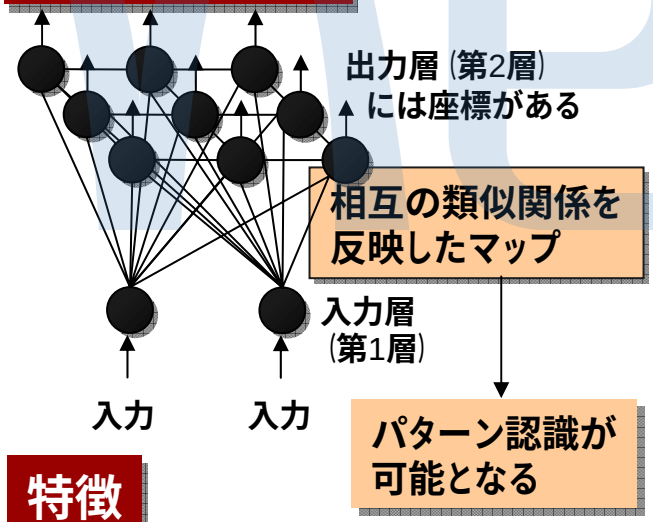
ホップフィールド法

MediBic 株式会社メディビック

自己組織化マップ– Self Organization Map

目的	分類や識別，予測，パターン認識
概要	ニューラルネットワークの一種 ・競合学習型ベクトル量子化ニューラルネットワーク ・多次元データの相互関係を低次元のマップで表示 ・相互関係は距離で表す

競合学習型NN



特徴

- 隠れ層や隠れユニットがない
- 出力層（第2層）のユニットは，1つ1つ　明確な位置（座標）を持つ

多変量解析分野におけるポジショニングマップとの違い

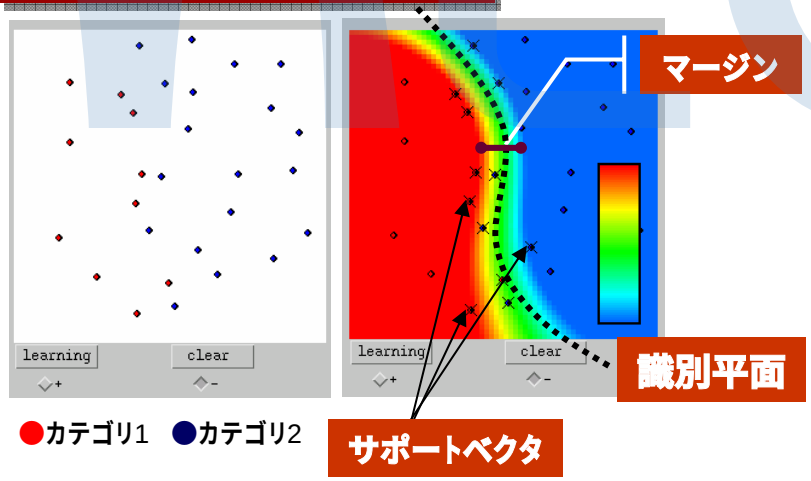
- ユークリッド距離空間で描かれていない
  - 1個離れたユニットとN個離れたユニットが同程度似ていないということあり得る
  - 「軸方向」の概念がない
- 多次元情報を2次元に折りたたんで表現できる
  i.e. マップを広くする ⇔ 相違を細かく表現できる
  ⇔ 対して，主成分分析では．．．
  - 2次元では，2つの主要特性しか考察できない
  - 表現できない情報は捨てられる

サポートベクタマシン– Support Vector Machine

目的	分類や識別，予測，パターン認識
概要	ニューラルネットワークの一種. ・単純パーセプトロンを用いて識別 ・識別機械の中でも識別能力が高いと言われている

カテゴリの境界を求めるために境界付近にあるデータ（サポートベクタ）を元に，マージン（幅のある境界）を求め，その中点を通る分離境界を決めるのが特徴

SVMによるデータ分類



マージンが最大になるようにサポートベクタを選択する

特徴

- **トレーニングデータからマージンを最大化するようにパラメータを学習**
- **過学習を起こしにくい**
- **様々なカーネルを利用可能**

http://mimi.aist-nara.ac.jp/~taku-y/private/applet/svm/

MediBic 株式会社メディビック

ナイーブベイズ– Naïve Bayes

目的	分類や識別
概要	統計的決定論 (*1) を用いた分類識別法 ・(過去の)集められたデータから推測し，データの数が多いほど確実性が　高くなる ・説明変数間の依存性を考慮

(*1) 数学者トーマス・ベイズが提唱した確率論で，過去の出来事の発生頻度を分析すれば未来の出来事を予想できるというもの．サーチエンジン大手のGoogleや情報検索ツールを販売するAutonomyの両社もベイズ理論を採用し，検索サービスを提供している．

Bayes Ruleの例

『未来の出来事の確率はその事象の過去の発生頻度を求めることで計算できる』

ある病気にかかっているか否かを判定する試薬を考える．試薬の反応は　＋(陽性)　or　-(陰性)．

- 対象となる病気にかかっている人の全人口に占める割合　：0.005
- 病気にかかっている人に対し試薬が陽性を示す確率　：0.98
- 病気にかかっていない人に対し試薬が陰性を示す確率　：0.97

この試薬に対して陽性の反応を示した場合，病気にかかっていると判断しますか?

P(disease)　=0.005　P(¬disease)　=0.995
P(+|disease)　=0.98　P(-|disease)　=0.02
P(-|¬disease)=0.97　P(+|¬disease)=0.03

Bayes Rule

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

P(disease|+) = 0.141

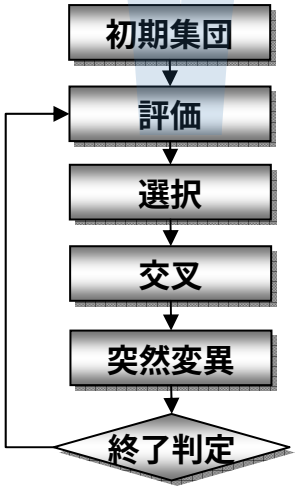
∴病気にかかっていると判断できない

遺伝的アルゴリズム– Genetic Algorithm

目的	最適化，学習，推論
概念	生物の進化の過程を真似ることで最適化を図る手法 ・遺伝と自然淘汰を繰り返すことで優秀なアルゴリズムを導き出す ・工学的には，最適解をランダムかつ速やかに探索する手法

環境により適応した個体がより高い確率で生き残り，次の世代に子を残すというメカニズムをモデル化し，環境に対して最もよく適応した個体，すなわち目的関数に対して最適値を与えるような解を計算機上で求めようというのが基本概念

遺伝的アルゴリズムフロー



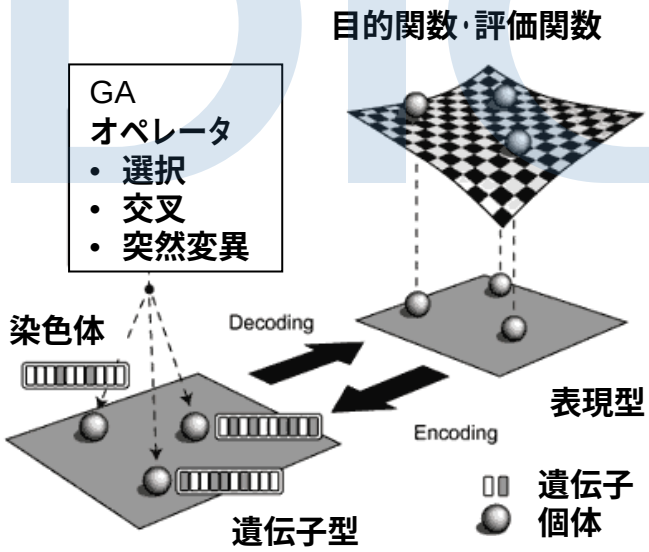
- ランダムに生成
- 各個体に評価値を与える
- 適者生存．適合度の高い個体．
- 有性生殖．新しい組み合わせ．
- 多様性

特徴

- 決定論的規則ではない
- 確率的オペレータを用いた探索
- 多点探索

問題点

- 目的に合わせた評価関数を都度設定しなければならない
- 計算負荷が高い
- 局所解へ収束することがある
- パラメータ設定が複雑である



D.E.Goldberg, "Genetic Algorithms in Search, Optimization and Machine Learning", Addison-Wasley Publishing Company, 1989

http://mikilab.doshisha.ac.jp/dia/research/person/jiro/reports/whatsGA/whatsGA01.html

MediBic 株式会社メディビック

【参考】早熟収束による局所解への収束