In Silico Data
Miracles by the KY-methods

◇第二部　化学多変量解析／パターン認識（ケモメトリックス（Chemometrics））関連(1)

Part2. Chemical multivariate analysis / pattern recognition (Chemometrics)

株式会社　インシリコデータ
湯田　浩太郎

Contents:

◆化学多変量解析／パターン認識（ケモメトリックス（Chemometrics））関連キーワード（1）

Chemical multivariate analysis / pattern recognition (Chemometrics) related keywords

化学パラメーター	Chemical parameters
2／3次元パラメーター	2/3 dimensional parameters
種々データ解析手法	Various data analysis methods
過剰適合	Overfitting
偶然相関	Chance correlation
線形／非線形性	Linear / Nonlinearity
特徴抽出	Feature selection,
最少サンプル数	Minimum number of samples
最少パラメーター数	Minimum number of parameters
クラスポピュレーション	Class population
次元変換／圧縮／縮小	Dimension conversion / compression / reduction
分類率／予測率	Classification rate / Prediction rate
要因抽出	Factor extraction

◆化学多変量解析／パターン認識（ケモメトリックス（Chemometrics））関連キーワード（2）

Chemical multivariate analysis / pattern recognition (Chemometrics) related keywords

オートスケーリング	Auto scaling
アウトライヤー／インライヤー	Outlier / Inlier
内挿／外挿問題	Interpolation / extrapolation problem
解析信頼性指標	Analysis reliability index
（サンプル数／パラメーター数）	(Number of samples / number of parameters)
クロスバリデーション	Cross validation
ROC／AUC	ROC／AUC
KY（K-step Yard sampling）法	KY (K-step Yard sampling) methods
パーセプトロン	perceptron
バックプロパゲーション	Back propagation
遺伝的アルゴリズム	Genetic algorithm
ブートストラップ	Bootstrap
アンサンブル学習	Ensemble learning
ファジー理論	Fuzzy theory
他	others

□データ解析の現状
Current status of data analysis

イントロのイントロ（１）: Introduction

抽選機（データ解析ソフトウエア）回したら玉（解）が出た
When the lottery machine (data analysis software) turns,
a ball (solution) comes out.

その玉（解）は、**当たり？　はずれ？**
Is that ball (solution) a hit? Are you off?

イントロのイントロ（２）: Introduction

◆ データ解析ソフトに**データを入れれば必ず答えが出る**

If you put data in the data analysis software, **you will always get an answer.**

⇒ これが**最も深刻な問題**です⇒ This is the **most serious problems**

勘違い: Misunderstandings

⇒ 解が出たから**プロフェッショナル**だー
I'm professional because I got the answer

⇒ 一件落着で仕事完了。**仕事したなーー・・・、疲れたけど達成感**
Completion of work with one calm. Feeling tired but feeling accomplished

⇒ 解を解析して、**次の仕事に向けた情報収集**しようか。
Would you like to analyze the solution and collect information for your next job?

＊そんなことして大丈夫かなーー？？ **解の信頼性**は？？
Is it okay to do that? ? How reliable is the solution??

⇒ データ解析手法を**数式レベルで展開**できるので間違うはずはない
Data analysis techniques can be expanded at the formula level so there should be no mistake

＊データ解析手法の**基本原理**と**運用**は**別問題**です
The basic principle and operation of data analysis methods are different issues

イントロのイントロ(3): Introduction

初心者:データ解析ソフトが動いて解が出てくるレベル

単にマニュアルに従って操作したら解が出た・・

＊素晴らしいですね、解が出てきたのですから。達成感あるでしょう・・・。

Beginner: Which data analysis software moves and solutions come out.
The solution came out simply by operating according to the manual ...

* It's wonderful, because the solution has come out. There will be a sense of accomplishment ...

イントロのイントロ(3): Introduction

質問①単にラッキーだっただけではありませんか？

・数値データにミッシングデータが混ざっていなかった

・数値データの桁数は？

多少桁数が変わっていてもプログラムは解を出します

・0データの扱いはどのようにしましたか？

・分類／予測率、相関／絶対係数しか気にしてませんか？

Question (1) Isn't she just lucky?

・ Missing data was not mixed with numerical data

・ Number of digits in numeric data? *

The program will give a solution even if the number of digits changes slightly.

・How did you handle the 0 data?

・Are you only interested in classification / prediction rate, correlation / absolute coefficient?

○ベテラン(データサイエンティスト):正しい答えを出す保証の有無

Veteran (data scientist): Is there a guarantee that the correct answer will be given?

○個々の分野の研究者:出てきた解を正しく評価／討論できる力

Individual field researchers: Ability to correctly evaluate / discuss solutions

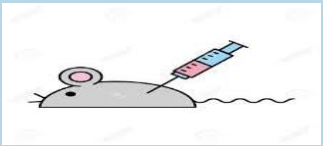
口本講演の目指すところ　Aim of this lecture

データ解析手法　基本原理等
Data analysis methods, basic principles, etc.

データ解析関連技術
Data analysis technology

道路交通法、マップ、天候
道路（坂、砂、高速、砂利等）
Road traffic law, map, weather
Road (slope, sand, high speed, gravel, etc.)

適用分野
Application fields



口本講演の目指すところ　Aim of this lecture

In Silico Data
Miracles by the KY-methods

データ解析手法　基本原理等
Data analysis methods, basic principles, etc.

二クラス分類
多クラス分類
重回帰
ニューラルネットワーク
バイナリーツリー
マッピング
クラスタリング
チャート分析
SVM
PCA
AdaBoost
ランダムフォレスト
その他

基本的な部分は説明
Basic part is explained

データ解析関連技術
Data analysis technology

○データ解析の適用限界
・個々の手法に依存
○サンプリング
・最小サンプル数
・クラスポピュレーション
・サンプリングプロトコル
・アウトライヤー、インライヤー
○線形／非線形問題
・空間合致と空間の再構築
・外挿と内挿
○パラメーター
・種類
・パラメーター選択
・0値の扱い
・スケーリング
○解析信頼性
・サンプル数／パラメーター数
・要因解析
・クロスバリデーション

本日の討論テーマ
Today's discussion theme

適用分野
Application fields

●創薬関連
・Q／T／ADME／P／SAR
・インシリコスクリーニング
・ドラグリストラクチャリング
・要因解析
●化合物毒性評価
・STR（構造毒性相関）
・毒性予測
・脱毒性
・化合物＆環境規制
●機能性化合物デザイン
・SPR（構造物性相関）
●機器スペクトル
・スペクトル解析
・メタボロミクス
●バイオ解析関連
●医療との連携（画像等）
●その他

適用事例
Case study

◇「ケモメトリックス」とは

ケモメトリックスとは：
計量化学（けいりょうかがく、chemometrics）とは、数理科学、統計学、機械学習、パターン認識、データマイニングなどの手法により、（広義の）化学分野における諸問題を解決しようとする分野である。

ウィキペディアより：https://ja.wikipedia.org/wiki/計量化学
＊ちなみに、chemometricsなので、「化学計量学」 by Yuta

Chemometrics is the science of extracting information from chemical systems by data-driven means. Chemometrics is inherently interdisciplinary, using methods frequently employed in core data-analytic disciplines such as multivariate statistics, applied mathematics, and computer science, in order to address problems in chemistry, biochemistry, medicine, biology and chemical engineering. In this way, it mirrors other interdisciplinary fields, such as psychometrics and econometrics.
https://en.wikipedia.org/wiki/Chemometrics

ケモメトリックスの適用研究分野
Application research field of chemometrics

多種多様な分野での研究の基本:Basics of research in various fields
（ケモメトリックスは様々な分野で適用できる汎用ツールである）

(Chemometrics is a general-purpose tool that can be applied in various fields)

構造－活性相関（QSAR）
構造－毒性相関（QSTR）
構造－物性相関（QSPR）
メタボロミクス　:Metabolomics
インシリコ（薬理活性/毒性/物性/ADME）スクリーニング
オフターゲット（off-target）創薬(Drug design)
マルチターゲット（multi-target）創薬
ドラグリポジショニング(Drug repositioning)
テーラーメードモデリング(Tailor-made modeling)
並列創薬(Parallel drug design)
インシリコンビ(Insilicombi)
医療解析(Medical analysis)
バイオ関連解析(Bio-related analysis)
その他(Others)

□ケモメトリックスによる解析実施の流れ
Flow of chemometrics analysis

In Silico Data
Miracles by the KY-methods

ケモメトリックス適用基本原理
Basic principles of chemometrics application

化合物情報
Compound information
ターゲット情報
Target information

多変量解析
パターン認識
人工知能
Multivariate analysis
Pattern recognition
Artificial intelligence

要因情報
新規デザイン
スクリーニング
Factor information
New design
screening

アナログ情報
Analog information

- 解析デザイン
- データ収集
- プロトコル検討
- その他
- Analysis design
- Data collection
- Protocol study
- Others

数値情報
Numerical data

- データ解析手法
- パラメーター
- 種々留意事項
- その他
- Data analysis method
- Parameter
- Various points to consider
- Others

アナログ情報
Analog information

- 解析ターゲット知識
- データの要因解析
- 外挿関連情報／知識
- その他
- Analysis target knowledge
- Data factor analysis
- Extrapolation-related information / knowledge
- Other

□ケモメトリックスの化学分野への適用原理
Principles of application of chemometrics to the chemical field

In Silico Data
Miracles by the KY-methods

種々データ解析でのケモメトリックス二大適用原理
Two major principles of chemometrics in various data analysis

◇ 宝箱アプローチ Treasure box approach

宝箱に相関やメカニズム等の重要な情報が隠されている

Important information such as correlation and mechanism is hidden in the treasure chest

＊ 仮設検証型アプローチ（既存アプローチ）

Temporary verification approach (existing approach)

解析前に、重要な情報に関する仮説等が必要。

Require hypotheses about important information before analysis.

◇ 発見型アプローチ Discovery approach

様々な重要情報を仮説抜きで発見（探索）する

Discover (search) various important information without hypothesis

＊ 仮設検証型アプローチ（既存アプローチ）

Temporary verification approach (existing approach)

解析前に仮設に関する何らかの情報や知識が必要

Require some information and knowledge about temporary structure before analysis

◇ 宝箱アプローチ：基本原理
Treasure box approach: Basic principles

多変量解析やパターン認識適用の基本原理
Basic principles of applying multivariate analysis and pattern recognition

情報等価原理
Information equivalence principle

化合物
蛋白
ゲノム

Compound protein genome

宝箱 Treasure box
（相関情報）
Accountability
(Correlation information)

薬理活性
毒性
副作用
代謝
分解性

Activity toxicity Side effects
Metabolism Degradability

情報等価性／必然性
Information equivalence / inevitable

原因情報
Target information

説明義務
Accountability

パターン認識手法
Pattern recognition

結果情報
Results

◇ 宝箱アプローチ：解析実施基本原理

Treasure box approach: Basic principle of analysis

多変量解析やパターン認識適用の基本原理

Basic principles of applying multivariate analysis and pattern recognition

情報等価原理

Information equivalence principle

◇発見型アプローチ：◇ Discovery approach

宝箱の中味は何だろう？

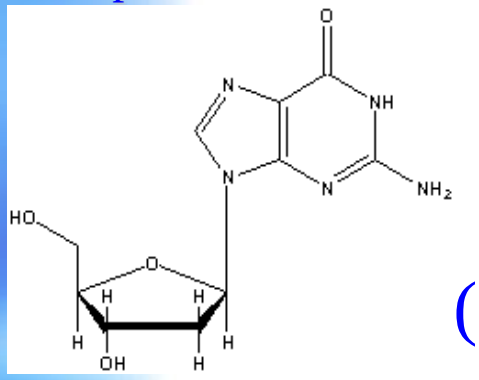
What's inside the treasure chest?

◇ 宝箱アプローチ：化合物関連分野

Treasure chest approach: Compound-related fields

化合物構造式

Compound structure



宝箱 Trasure box

（相関情報）

(Correlation information)

薬理活性
毒性
副作用
代謝
分解性

Activity toxicity Side effects
Metabolism Degradability

情報の等価性／必然性

Information equivalence / inevitable

種々パラメータ発生

Various parameter generation

初期パラメータ群
分子量
原子／結合数
HOMO／LUMO
その他

Initial parameter group
Molecular weight
Number of atoms / bonds
HOMO / LUMO
Other

特徴抽出

Feature selection

最終
パラメータ群
Final
parameter set

種々の解析手法
・判別分析
・クラスタリング
・フィッティング
・その他

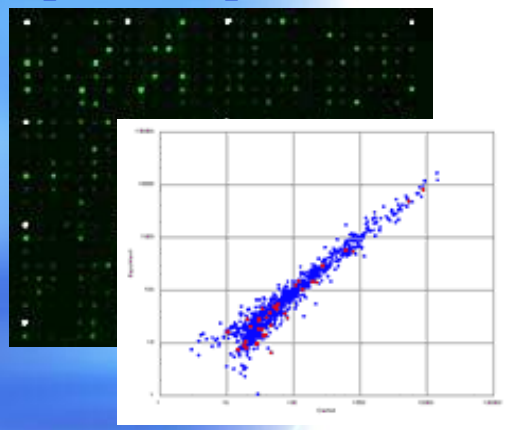
・Discriminant analysis
・Clustering・Fitting・Others

◇ 宝箱アプローチ：バイオ関連分野

Treasure box approach: Bio-related fields

発現プロフィールデータ

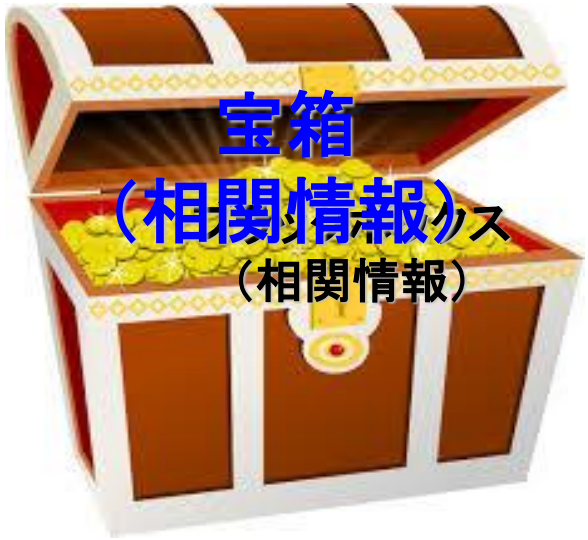
Expression profile data



数値パラメータ

Numerical parameters

Hs.35092 ;5523
Hs.27935 ; 857
Hs.622 ;3682
Hs.38677 ;2283
・ ・ ・ ・ ・ ・



種々の解析手法

Various data analysis methods

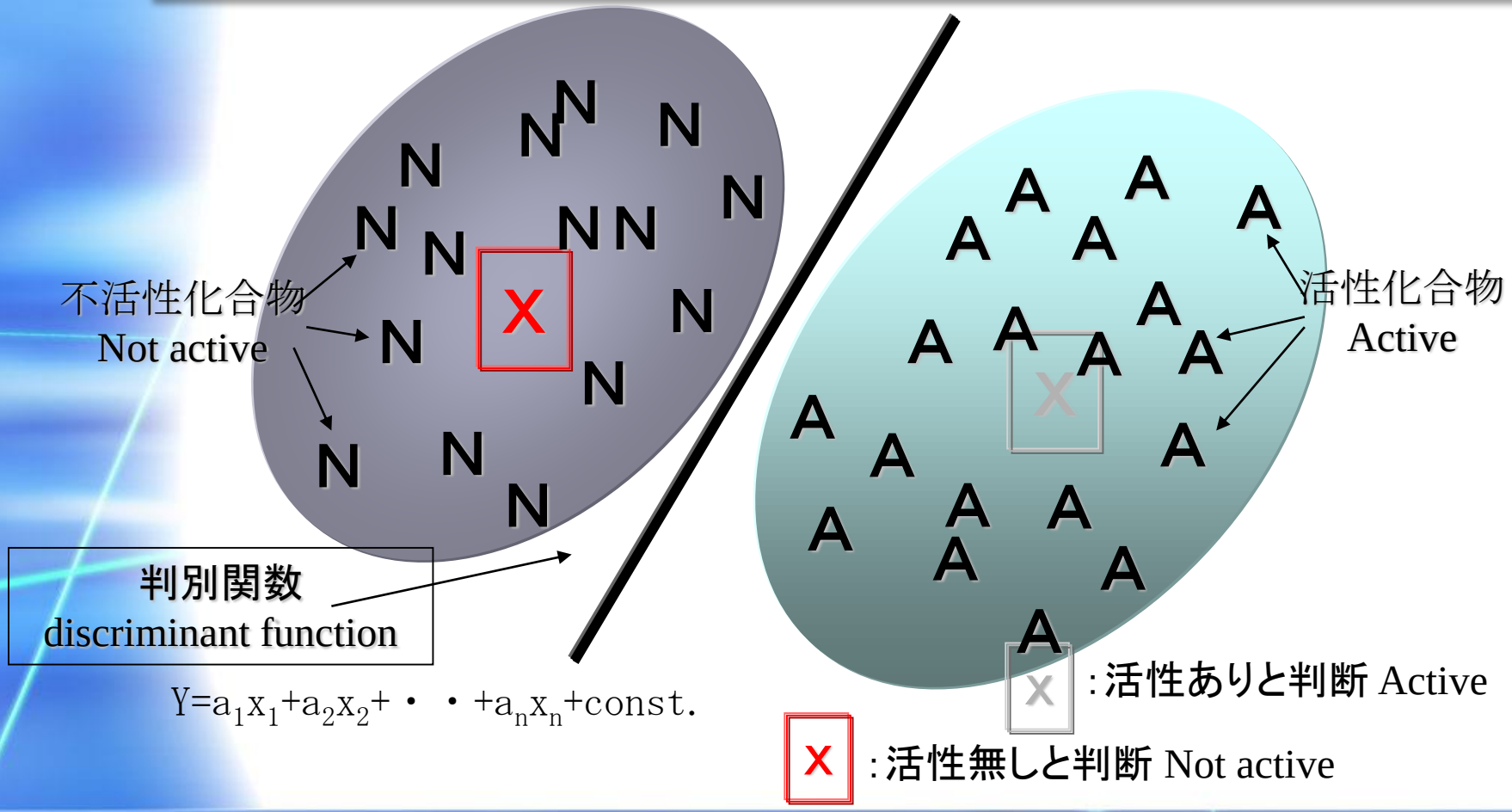
- 判別分析
- クラスタリング
- マッピング
- フィッティング
- その他

薬理活性
毒性
副作用
代謝
分解性

Activity toxicity Side effects
Metabolism Degradability

□二クラス分類の基本概念(1)
□ Basic concept of binary classification

判別関数を用いた化合物の分類／予測のイメージ
Image of compound classification / prediction using discriminant function



□二クラス分類の基本概念(2)
Basic concept of binary classification

判別関数とドラグデザイン的情報解析
Discriminant function and drag design information analysis

$$Y = a_1x_1 \pm a_2x_2 \pm \cdots \pm a_nx_n \pm \text{const.}$$

Y ： 目的薬理活性／毒性／他
Activity, toxicity, others

Y ≧ 0
・活性あり Active
・毒性あり Toxic

Y ＜ 0
・活性無し Not active
・毒性無し No toxic

□構造－活性／毒性／他との相関解析
Structure-activity / toxicity / correlation analysis with others

係数 Coefficient $a_i \geqq 0$
パラメータXiの持つ情報は
・活性向上、・毒性強化要因
Information of parameter Xi is:
・Activity improvement ・Toxicity enhancement factor

係数 Coefficient $a_i < 0$
パラメータXiの持つ情報は
・活性低下、毒性低下要因
The information of parameter Xi is:
・Activity reduction, toxicity reduction factor

構造－活性相関、構造－毒性相関
Structure-activity relationship, structure-toxicity relationship

二クラス分類の基本概念(3)
Basic concept of binary classification

判別関数、重回帰式からの情報取り出し
Extract information from discriminant function and multiple regression equation

パラメータとウェイトベクトルの利用 Using parameters and weight vectors

1. 要因抽出および明確化：パラメータ利用Factor extraction and clarification: use of parameters
   最終判別関数や重回帰式中で用いられているパラメータが持つ情報内容の評価
   Evaluation of information content of parameters used in final discriminant functions and multiple regression equations

2. 寄与の方向性：ウエイトベクトル利用 Direction of contribution: Using weight vectors

①係数 Coefficient $a_i \geqq 0$ ⟹ パラメータXiの情報はThe information of parameter Xi is
・活性向上、あるいは毒性強化要因となる・Activity improvement or toxicity enhancement factor

②係数 Coefficient $a_i < 0$ ⟹ パラメータXiの持つ情報は
・活性低下、毒性低下要因・Activity reduction, toxicity reduction factor

3. 寄与の相対的な貢献度：ウエイトベクトル利用 Relative contribution of contribution: use of weight vector
   係数の絶対値比較による貢献度の評価
   Evaluation of contribution by comparing absolute values of coefficients

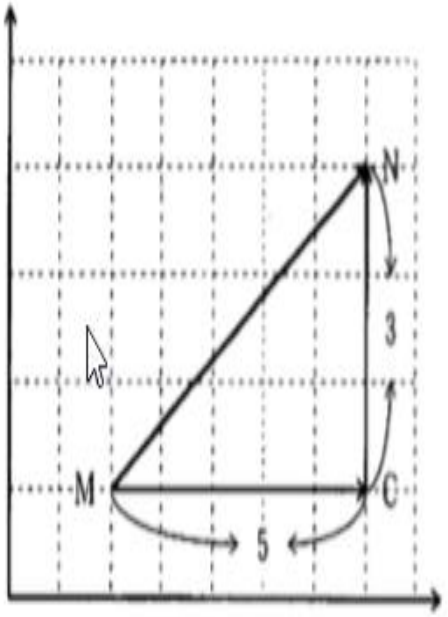
□多変量解析／パターン認識の基本事項
Basics of multivariate analysis / pattern recognition

◇サンプル間の**距離** Distance between samples

N次元空間上でのサンプル間の距離の算出は様々な多変量解析／パターン認識手法の基本となる。Calculation of the distance between samples in the N-dimensional space is the basis of various multivariate analysis / pattern recognition techniques.

サンプル間の距離算出に様々なメトリック手法が展開されている。Various metric techniques have been developed for calculating the distance between samples.

右は、ユークリッド距離とシテイブロック距離による距離計算事例である
On the right is an example of distance calculation using Euclidean distance and city block distance.



2次元空間中でA点とB点との距離を計る時、ユークリッド距離では距離D_{MN}を算出し、シテイブロック距離では距離D_{MC}と距離D_{CN}との距離を合わせたものをA点とB点との距離とする。

M＝（2，1）
N＝（7，4）

ユークリッド距離 ＝ $D_{MN} = (5^2 + 3^2)^{1/2} = 5.83$

シテイブロック距離 ＝ $D_{MC} + D_{CN} = 5 + 3 = 8$

図1．　ユークリッド距離とシテイブロック距離

◇サンプル間の距離：**距離基準**

Distance between samples: Distance standard

(1) ミンコフスキー (MINKOWSKI)距離

$$D = \left[\sum_{i=1}^{d} (X_{Mi} - X_{Ni})^{k} \right]^{1/k}$$

(2) ユークリッド(EUCLIDEAN) 距離

ユークリッド距離はミンコフスキー距離の式において、kが2の時にあたる。

$$D = \left[\sum_{i=1}^{d} (X_{Mi} - X_{Ni})^{2} \right]^{1/2}$$

(3) シテイブロック(CITY BLOCK)距離

シテイブロック距離は2パターン間の最短距離をとるのではなく、直交する2線の距離の総和をとるものである。

$$D = \sum_{i=1}^{d} \left[X_{Mi} - X_{Ni} \right]$$

(4) キャンベラ (CANBERA)距離

$$D = \frac{\sum_{i=1}^{d} \left[X_{Mi} - X_{Ni} \right]}{\sum_{i=1}^{d} \left[X_{Mi} + X_{Ni} \right]}$$

(5) ハミング(HAMMING) 距離

ハミング距離は1／0のバイナリデータで利用される事が多いが、ORとANDの識別を効率良く行う事が出来る。

$$D = \sum_{i=1}^{d} \left[X_{Mi} + X_{Ni} - 2X_{Mi}X_{Ni} \right]$$

尚、1／0のバイナリーデータを用いた時、ハミング距離とシテイブロック距離は同じものとなる。

(6) 谷本距離

谷本距離はハミング距離がその性格上、1が少ないデータは不利に評価されるという欠点を改良したものである。

$$D = \frac{\sum_{i=1}^{d} \left[X_{Mi} + X_{Ni} - 2X_{Mi}X_{Ni} \right]}{\sum_{i=1}^{d} \left[X_{Mi} + X_{Ni} - X_{Mi}X_{Ni} \right]}$$

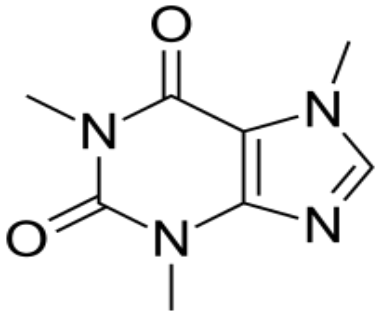
□化学データ（パラメーター）の種類
Chemical data (parameter) types

化合物に起因するデータを「化学データ（パラメーター）」とする
構造式の有無や内容により、以下の3種類に分類される

Three types of chemical parameters

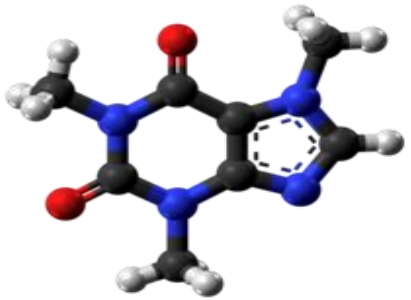
構造式不要
No structure

物性データ
機器スペクトルデータ
医療関連データ
Physical property data
Instrument spectral data
Medical data



2次元構造式関連データ
Two-dimensional structural formula related data

分子式関連パラメーター
トポロジカルパラメーター
部分構造パラメーター
フラグメントパラメーター
Molecular formula related parameters
Topological parameters
Substructure parameter
Fragment parameter



3次元構造式関連データ
Three-dimensional structural formula related data

トポグラフィカルパラメーター
3次元構造関連データ
力場関連パラメーター
量子化学関連パラメーター
Topographic parameters
3D structure related data
Force field related parameters
Quantum chemistry-related parameters

□化学データ(パラメーター)の種類
Chemical data (parameter) types

◇化合物に由来する化学データ（パラメーター）
Chemical data (parameters) derived from compounds

化合物構造式由来のパラメーター
Parameters derived from compound structural formula

■ トポロジカルデータ
分子構造インデックス：原子数（原子種）、結合数（結合種）、リング数、その他
様々なインデックス値：HOSOYAインデックス、分子結合インデックスMC値
パス値インデックス、

■ トポグラフィカルデータ
化合物の３次元的形状に関するパラメータ
化合物全体構造　：ボックスパラメータ、対称パラメータ、
立体格子パラメータ、その他
化合物部分構造　：ステリモルパラメータ、

■ 物理化学データ
分子に関する様々な物性データ　：　分子屈折率、分子量、LOGP、融点、沸点
分子容積、分子表面積、その他
分子軌道法より得られる様々なパラメータ：電子密度、HOMO、LUMO、他
分子力学計算から得られるパラメータ：　種々歪みエネルギー
種々スペクトルより得られるデータ：　種々スペクトルデータ

■ その他のデータ
部分構造パラメータ：　部分構造の有無、部分構造数、
部分構造単位の様々なパラメータ値計算、
演算パラメータ１　：　記述子間の演算により得られるパラメータ（＋－ｘ÷ Log）
演算パラメータ２　：　他の解析手法より算出されたパラメータ
ダミーパラメータ　：　有るパターン存在の有無（１／０）に関するパラメータ

機器スペクトル
Instrument spectrum

Mass
IR
H-NMR
C-NMR
UV
GC
HPLC
Raman
X線分析

その他

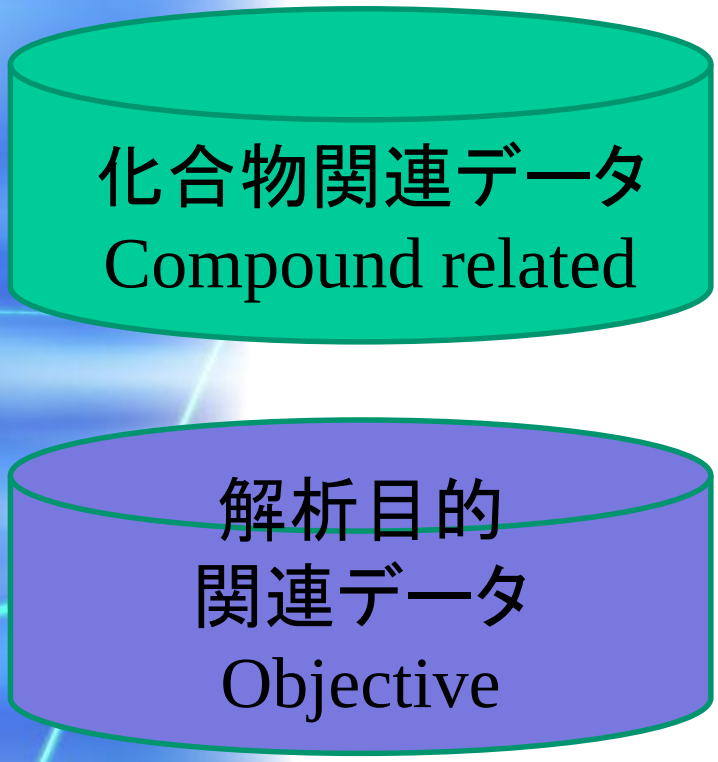
化合物に起因する解析目的パラメーター
Objective

薬理活性
毒性
ADME
物性
環境毒性

Pharmacological activity
toxicity
ADME
Physical properties
Environmental toxicity

□化学データ解析での必要データ
Required data for chemical data analysis

◇汎用的なデータ解析の流れ図：入力データ関連
General data analysis flowchart: Input data related



＊化合物構造式（1／2次元）
＊物性データ
＊機器スペクトルデータ
＊バイオ関連データ
＊医療関連データ（画像等）

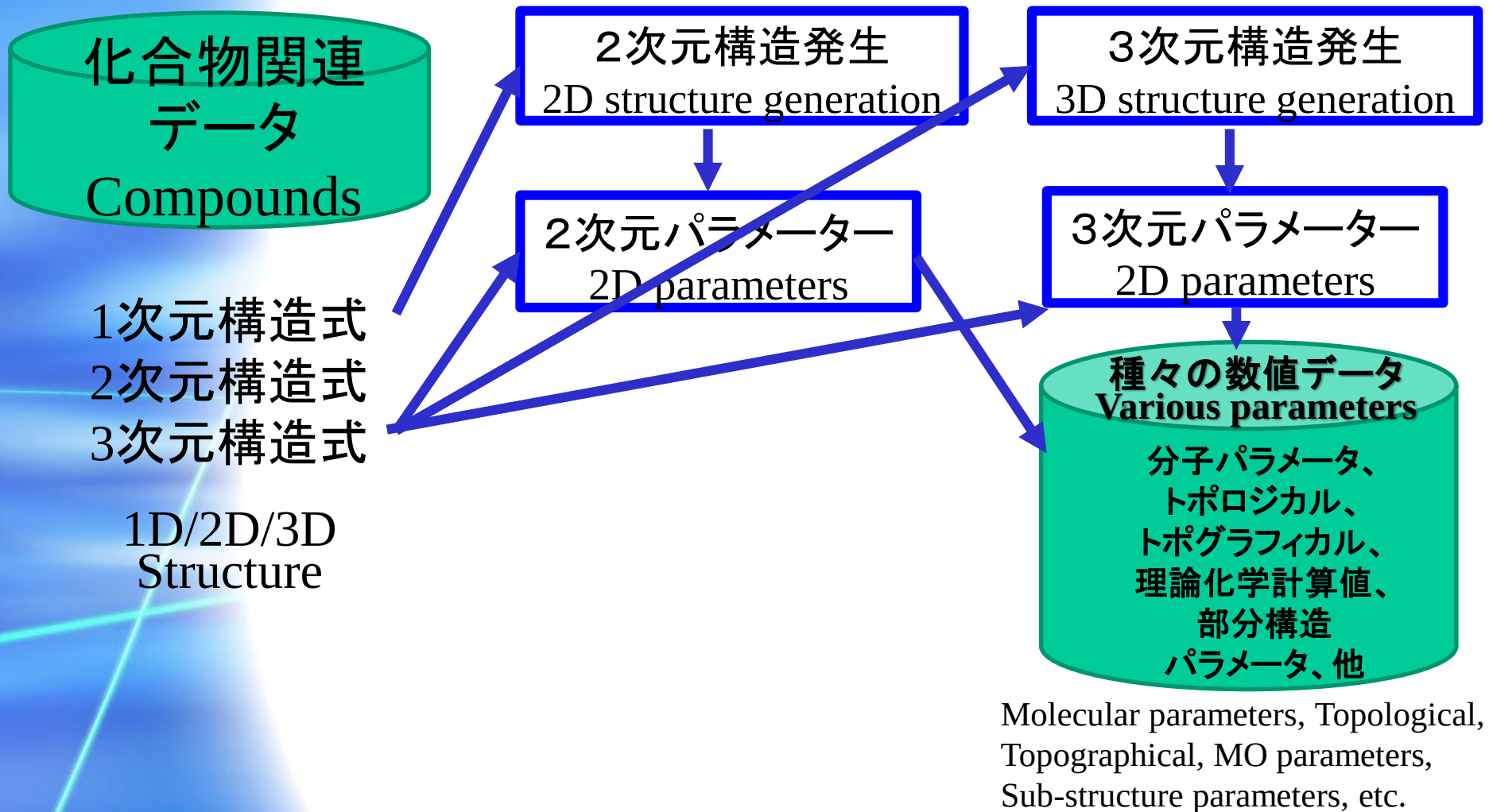
* Compound structural formula (1/2 dimension)
* Physical property data,* Instrument spectral data
* Bio-related data,* Medical related data (images, etc.)

＊薬理活性
＊毒性関連データ
＊機器スペクトルデータ
＊バイオ関連データ
＊患者関連データ

* Pharmacological activity,* Toxicity-related data
* Instrument spectral data,* Bio-related data
* Patient related data

◇化合物構造式を用いたパラメーター発生

Parameter generation using compound structural formula



Molecular parameters, Topological, Topographical, MO parameters, Sub-structure parameters, etc.

◇汎用的なデータ解析の流れ図：データ前処理関連
General data analysis flowchart: Data preprocessing

解析目的と無関係な情報を有するパラメーター除去（特徴抽出）
Parameter removal (feature selection) with information unrelated to analysis purpose

種々の数値データ
Various parameters

分子パラメータ、
トポロジカル、
トポグラフィカル、
理論化学計算値、
部分構造パラメータ、
演算パラメーター
その他

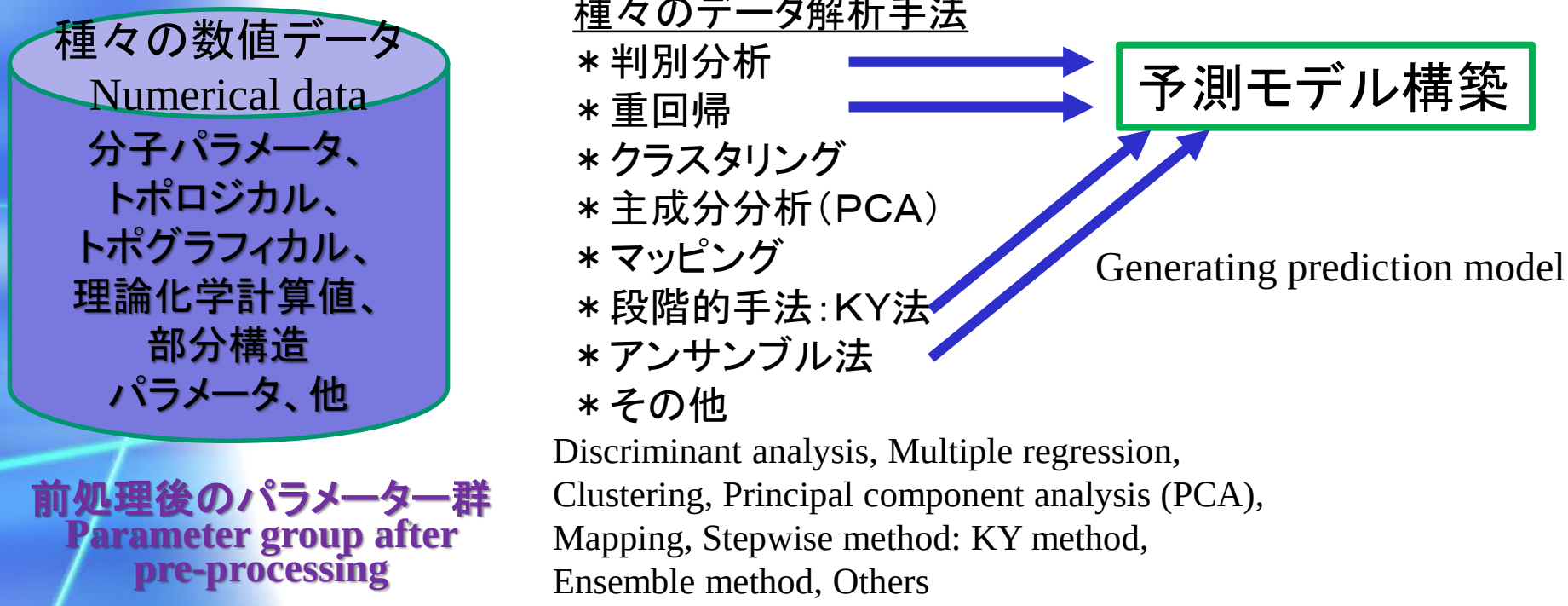
Molecular parameters, topological, Topographical, MO calculations, Substructure parameters, Calculation parameters, Other

パラメーター選択（特徴抽出）の実施
Implementation of parameter selection (feature selection)

＊欠陥データ処理：Missing data
　欠損データ、0値データ、同値データ、他
＊統計的処理：Statistical processing
　単相関、重相関、Fisher比、他
＊データ解析手法による個別特徴抽出
　Data analysis depended feature selection
　主成分解析、パーセプトロン、遺伝的アルゴリズム、他
＊パラメーター桁の調整 Digit adjustment
　オートスケーリング Auto scaling

◇汎用的なデータ解析の流れ図：多変量解析／パターン認識関連

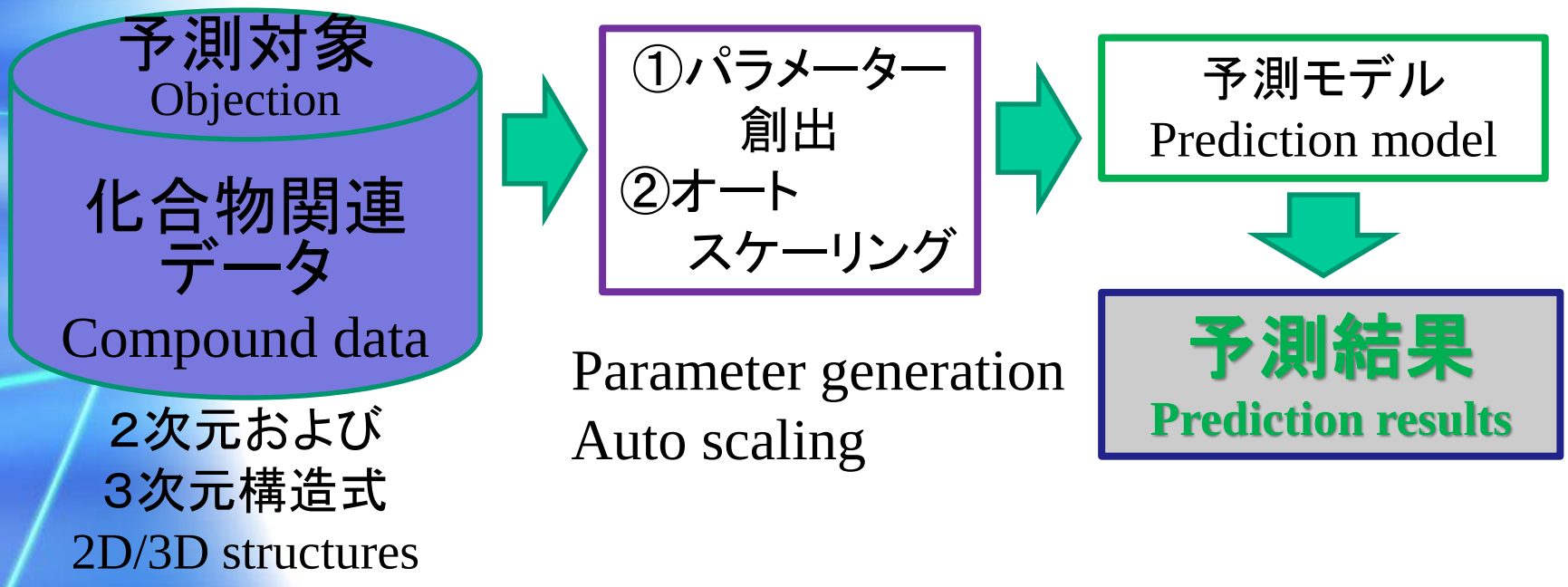
Flow chart of general-purpose data analysis: Multivariate analysis / pattern recognition



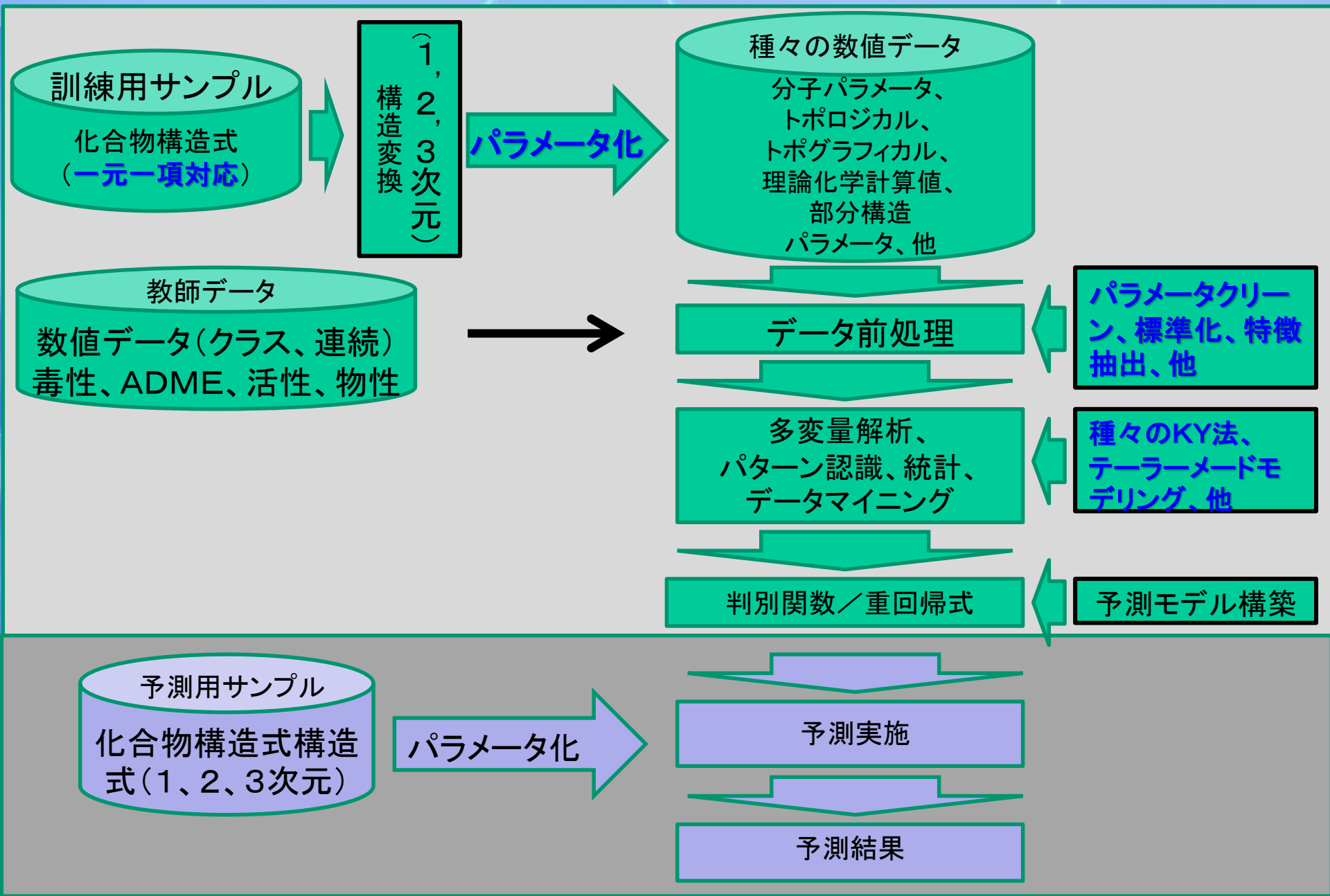
◇汎用的なデータ解析の流れ図：予測の実施

Flow chart of general-purpose data analysis: generate prediction models

＊予測モデルを構成するパラメーターを化合物構造式から創出
Create parameters that make up the prediction model from compound structural formulas
＊構造式から発生できないパラメーターは外部パラメーターとして導入
Parameters that cannot be generated from the structural formula are introduced as external parameters.
＊オートスケーリングを実施: Autoscaling



◇汎用的なデータ解析の流れ図：解析全体の流れ
Flow chart of general-purpose data analysis:data analysis flow



データ解析手法の簡単な紹介

A brief introduction to data analysis techniques

□ データ解析手法の機能的分類
Functional classification of data analysis methods

統計関連情報の扱い方によりデータ解析手法は大きく二種類に分類される

1. パラメトリック手法 Parametric methods

解析に用いるパラメーターの母集団分布情報を用いて解析する手法
分布（正規分布等）等の情報、平均値、分散、サンプル数は大きいものが良い
統計の検定等を実施するときに利用される

Analysis method using population distribution information of parameters used for analysis.
· Information such as distribution (normal distribution, etc.), average value, variance, number of samples should be large.
Used when conducting statistical tests.

2. ノンパラメトリック手法 Nonparametric method

解析に用いるパラメーターの母集団分布情報が不明な環境で解析する手法
分布型は問わない、サンプル数も小さくても実施可能、他
制限事項が少ないので、広範囲にわたって適用できる
基本的に多変量解析／パターン認識として展開される手法はノンパラメトリック

Analysis method in an environment where the population distribution information of parameters used for analysis is unknown.
Any distribution type is possible, even if the number of samples is small, etc.
Since there are few restrictions, it can be applied over a wide range.
Basically, the method developed as multivariate analysis / pattern recognition is nonparametric

◇ データ解析手法の機能的分類
Functional classification of data analysis methods

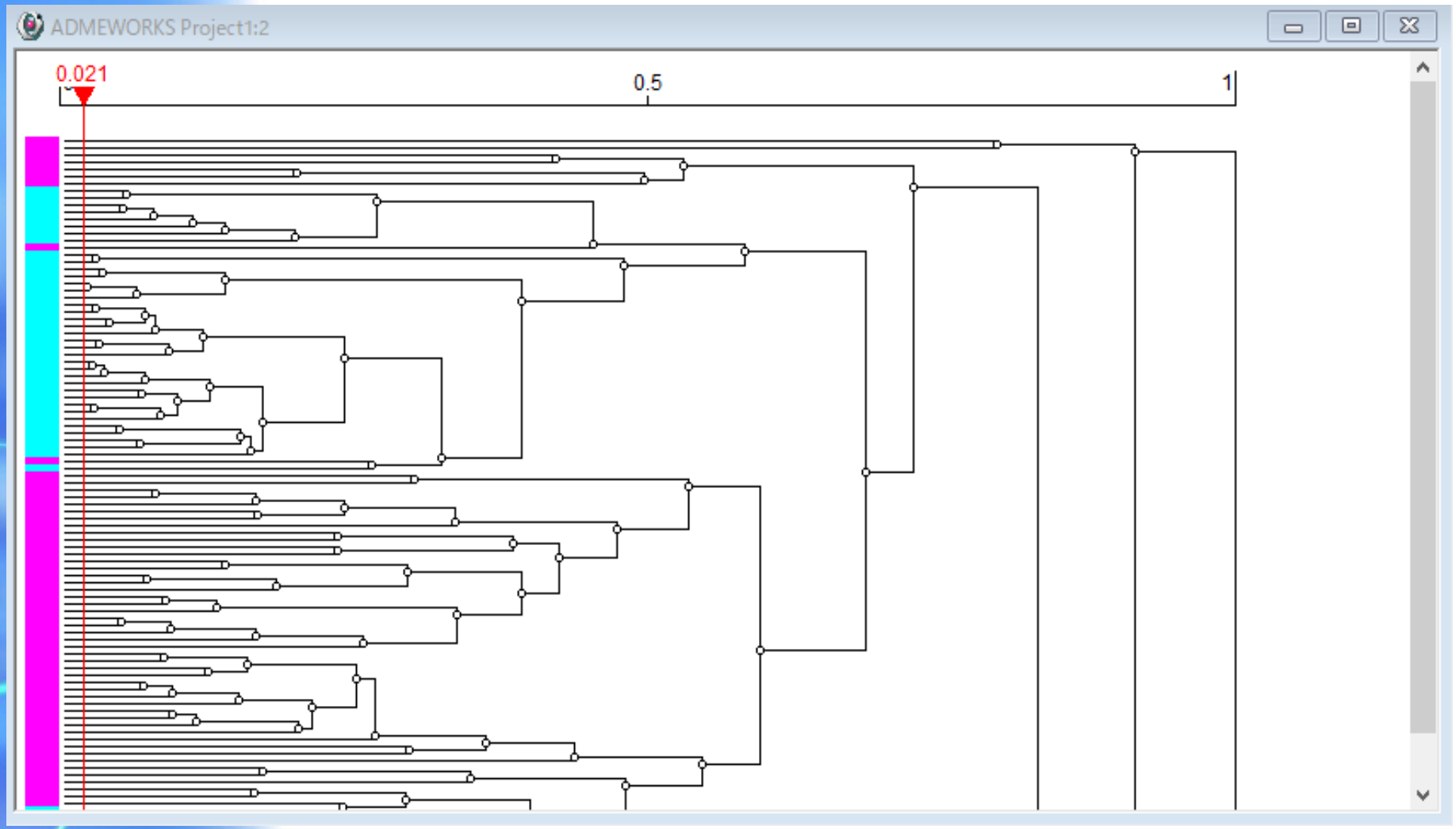
解析アプローチの差異に伴い大きく以下の解析手法に分類される
個々の手法での代表的な手法をリストする

1. 判別分析：二クラス分類、多クラス分類
2. 重回帰（フィッテイング）：単回帰、重回帰
3. クラスター分析：階層型クラスタリング、非階層型クラスタリング
4. 種々マッピング：主成分分析（PCA）、非線形写像（NLM）
5. チャート表示：手法⇒レーダーチャート、顔チャート、ラインチャート、
6. 決定木：手法⇒C5.0、CART
7. アンサンブル学習法：手法⇒AdaBoost、ランダムフォレスト
8. 段階的手法：手法⇒KY法
9. 最適化手法：最小二乗法、シンプレクス法、遺伝的アルゴリズム

1. Discriminant analysis: two-class classification, multi-class classification
2. Multiple regression (fitting): single regression, multiple regression
3. Cluster analysis: Hierarchical clustering, non-hierarchical clustering
4). Various mapping: principal component analysis (PCA), nonlinear mapping (NLM)
5. Chart display: method ⇒ radar chart, face chart, line chart,
6). Decision tree: Method ⇒ C5.0, CART
7). Ensemble learning method: Method ⇒ AdaBoost, Random Forest
8). Stepwise method: Method ⇒ KY method
9. Optimization method: least square method, simplex method, genetic algorithm

◇個別手法の特徴と適用内容：クラスター分析

Characteristics and application contents of individual methods: Cluster analysis



ModelBuilderの画面より

◇個別手法の特徴と適用内容：クラスター分析

Characteristics and application contents of individual methods: Cluster analysis

クラスター分析は以下の基準にて様々な手法に分類される

Cluster analysis is classified into various methods according to the following criteria:

クラスター化アプローチ

■階層的クラスタリング

Hierarchical clustering

解析結果はデンドログラムとして出力される

■非階層的クラスタリング

Non-hierarchical clustering

解析結果は単にクラスターの数とメンバー

クラスター化アルゴリズム

①Division Method（分割法）

② Aggregative Method（凝集法）

融合法の種類

① 最近隣法 (NEAREST NEIGHBOR METHOD)
② 最遠隣法 (FURTHEST NEIGHBOR METHOD)
③ 群平均法 (GROUP-AVERAGE METHOD)
④ 重心法 (CENTROID METHOD)
⑤ メジアン法 (MEDIAN METHOD)
⑥ ワード法 (WARD METHOD)
⑦ 可変法 (FLEXIBLE METHOD)

◇個別手法の特徴と適用内容：クラスター分析

Characteristics and application contents of individual methods: Cluster analysis

クラスタリングの融合手法　　Clustering fusion method

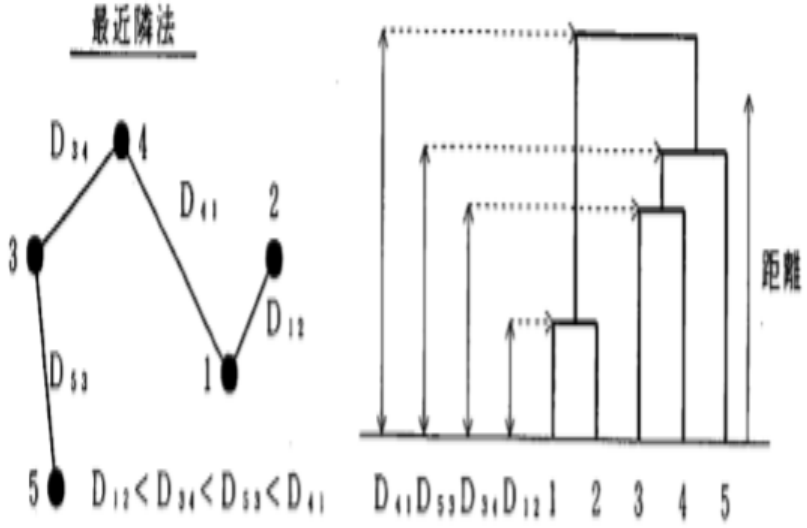


図2、　最近隣法によるクラスタリング手続きとデンドログラム

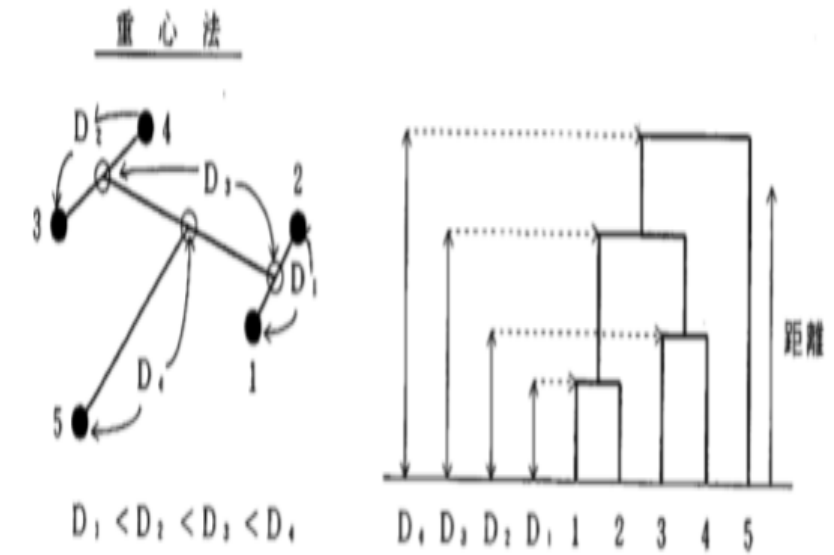


図3．重心法によるクラスタリング手続きとデンドログラム

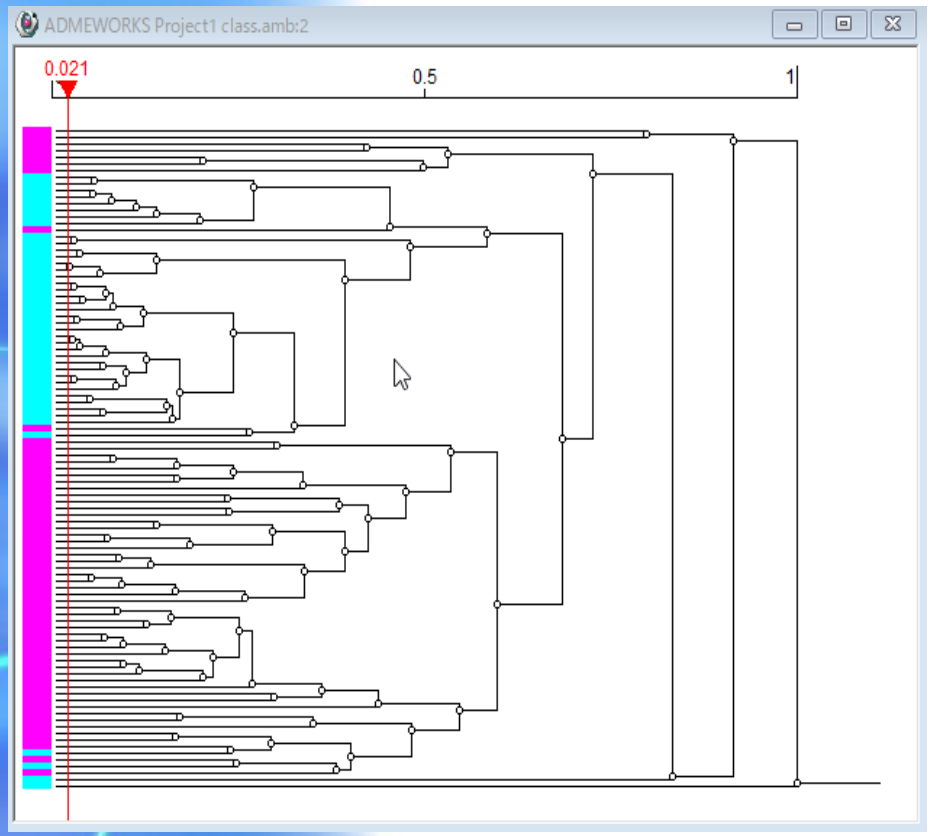
ModelBuilderの画面より

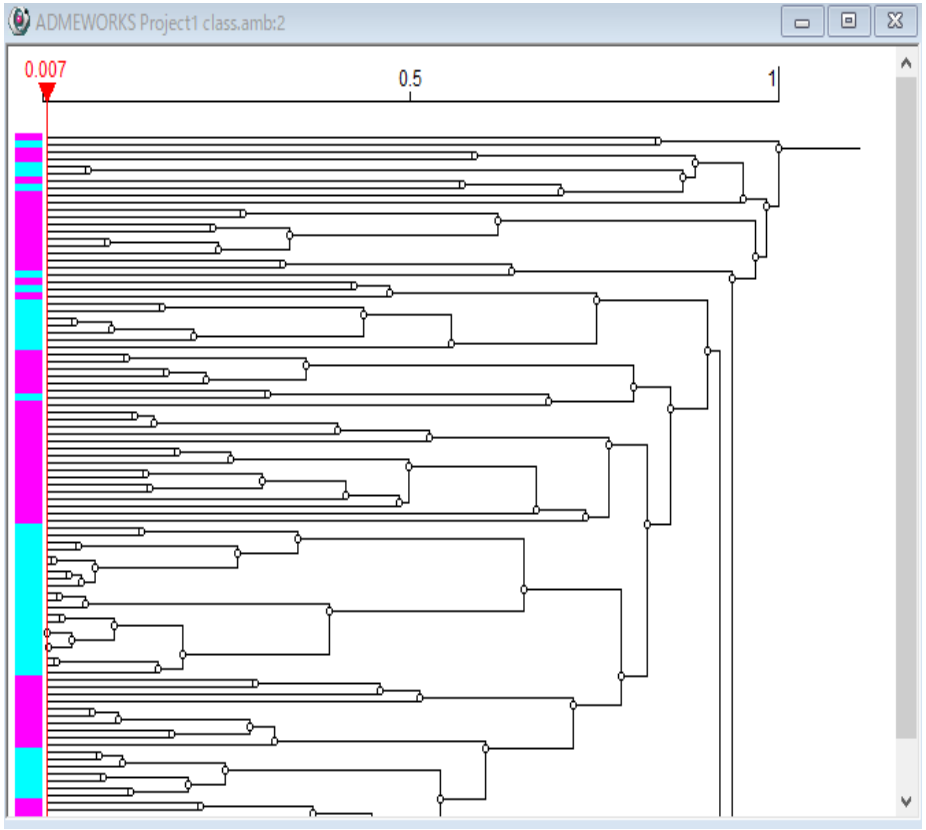
◇個別手法の特徴と適用内容：クラスター分析

Characteristics and application contents of individual methods: Cluster analysis

＊クラスタリングはサンプル同士の相対的な位置関係や近隣関係を俯瞰して見ることが出来る
Clustering provides a bird's-eye view of relative positional relationships and neighborhood relationships between samples.

＊メトリックや融合手法の差異によりクラスタリング結果が大きく変化するので要因解析は注意
Because the clustering results change greatly due to differences in metrics and fusion methods, be cautious of factor analysis





ModelBuilderの画面より

◇個別手法の特徴と適用内容：マッピング
Characteristics and application contents of individual methods: Mapping

■主成分分析（PCA）
- ・データ解析的にはリニアプロジェクション手法
- ・サンプル空間中のサンプルの位置関係を変えない
- ・オリジナルのサンプル空間を俯瞰する方向性を変えて空間を見直す手法
- ・PCA適用前と後とで次元（パラメーター）数の変化はない

■ Principal component analysis (PCA)
- ・Linear projection method for data analysis
- ・Do not change the positional relationship of the sample in the sample space
- ・Change the direction of bird's-eye view of the original sample space and review the space
- ・There is no change in the number of dimensions (parameters) before and after applying PCA

■非線形写像（NLM）
- ・データ解析的にはマッピング（写像）手法
- ・サンプルを人間が可視できる二次元/三次元上に分散
- ・サンプル空間を強制的に2/3次元に圧縮する
- ・非線形写像実施後は、サンプル空間の次元は2/3次元となる

■ Nonlinear mapping (NLM)
- ・For data analysis, mapping method
- ・Distribute samples in 2D / 3D that humans can see
- ・Forcibly compress the sample space into 2/3 dimensions
- ・After implementation of non-linear mapping, the dimension of the sample space is 2/3.

◇個別手法の特徴と適用内容：NLM

Characteristics and application contents of individual methods: Non linear mapping

◇個別手法の特徴と適用内容：Non Linear Mapping（NLM）

＊NLMにより、最初のN次元空間が可視できる2次元空間に変換された。

By NLM, the first N-dimensional space was converted into a visible two-dimensional space.

**＊NLMでは、元のN次元空間における
サンプル間の位置関係を保ちつつ
新しい2次元空間上に再配置される。**

＊ In NLM, the sample is rearranged in a new two-dimensional space while maintaining the positional relationship between samples in the original N-dimensional space.

**＊2次元空間上の第6サンプルは、
その他のサンプル（1～5、7～9）との
相互位置関係（空間上の距離）をN次
元空間上における関係と略同じとなる。**

＊ The 6th sample in the 2D space has the same mutual positional relationship (distance in space) with the other samples (1-5, 7-9) as the relationship in the N-dimensional space.

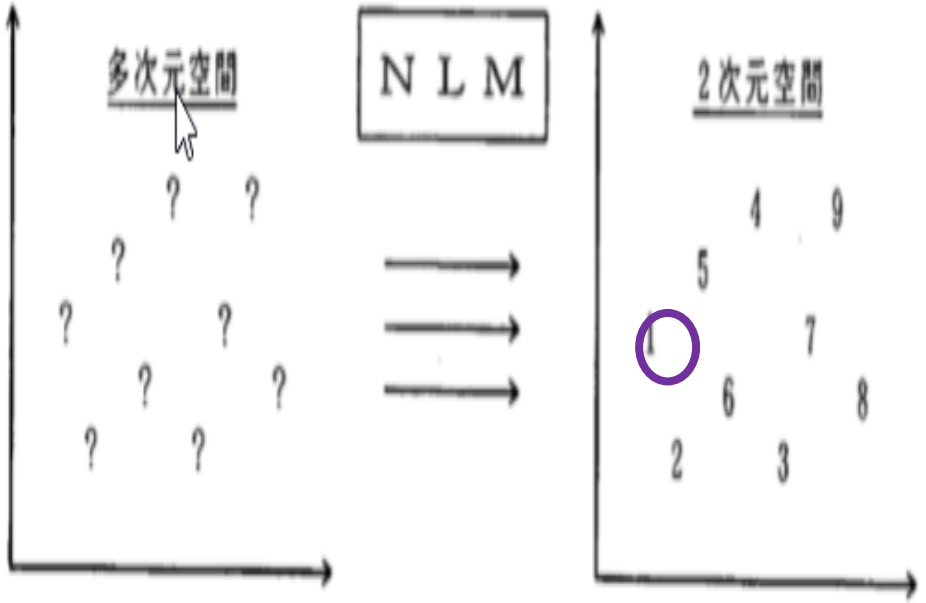


図5－6　ノンリニアマッピングによる多次元空間の2次元空間への写像

◇個別手法の特徴と適用内容：NLM

Characteristics and application contents of individual methods: Non linear mapping

◇個別手法の特徴と適用内容：Non　Linear　Mapping（NLM）

多次元（d次元）空間上のパターンA，B

$$PA = (X_{A1},\ X_{A2},\ X_{A3},\ \cdots\cdots,\ X_{A,d-1},\ X_{A,d})$$

$$PB = (X_{B1},\ X_{B2},\ X_{B3},\ \cdots\cdots,\ X_{B,d-1},\ X_{B,d})$$

2次元空間上のパターンM，N

$$P2A = (Y_{A1},\ Y_{A2})$$

$$P2B = (Y_{B1},\ Y_{B2})$$

$$DX_{AB} = \sum_{i=1}^{d} \{(X_{Ai} - X_{Bi})^2\}^{1/2} \quad \text{——————} \quad (\ )$$

$$D2_{AB} = \sum_{i=1}^{2} \{(Y_{Ai} - Y_{Bi})^2\}^{1/2} \quad \text{——————} \quad (\ )$$

$$Ei = \frac{1}{\sum_{A\neq B}^{n} DX_{AB}} \sum_{A\neq B}^{n} \frac{(DX_{AB} - D2_{AB})^2}{DX_{AB}} \qquad (i = 1,\ 2,\ \sim,\ n)$$

ここで、nはパターンの数を示す。

このエラー関数の極小化は（Y_{i1}，Y_{i2}）（i＝1～n）の初期チャートを適当に与え

$$\frac{\partial E(Y_{ij})}{\partial Y_{ij}} = 0 \quad (i = 1 \sim n;\ \ j = 1,\ 2) \quad \text{——————} \quad (\ )$$

$$Y_{ij}(m+1) = Y_{ij}(m) - k\Delta_{ij}(m) \quad \text{——————} \quad (\ )$$

ここで $\Delta_{ij}(m) = \dfrac{\partial E(m)}{\partial Y_{ij}(m)} \Big/ \left| \dfrac{\partial^2 E(m)}{\partial Y_{ij}(m)^2} \right|$ で、$k \fallingdotseq 0.3 \sim 0.4$

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

ニューラルネットワークによる次元圧縮
Dimensional compression by neural network

＊入力層のN個のパラメーターから中間層のユニット数（図は3）に次元を変換する手法＊ Method to convert dimensions from N parameters in the input layer to the number of units in the intermediate layer (3 in the figure)

＊NLMと異なり、サンプル間の相互位置関係はキープされない
＊ Unlike NLM, the mutual positional relationship between samples is not kept.

＊新たな次元は、ニューラルネットワークが解析目標とするクラスデータ等を説明する情報を含んでいる＊
The new dimension includes information that explains the class data etc. targeted by the neural network.

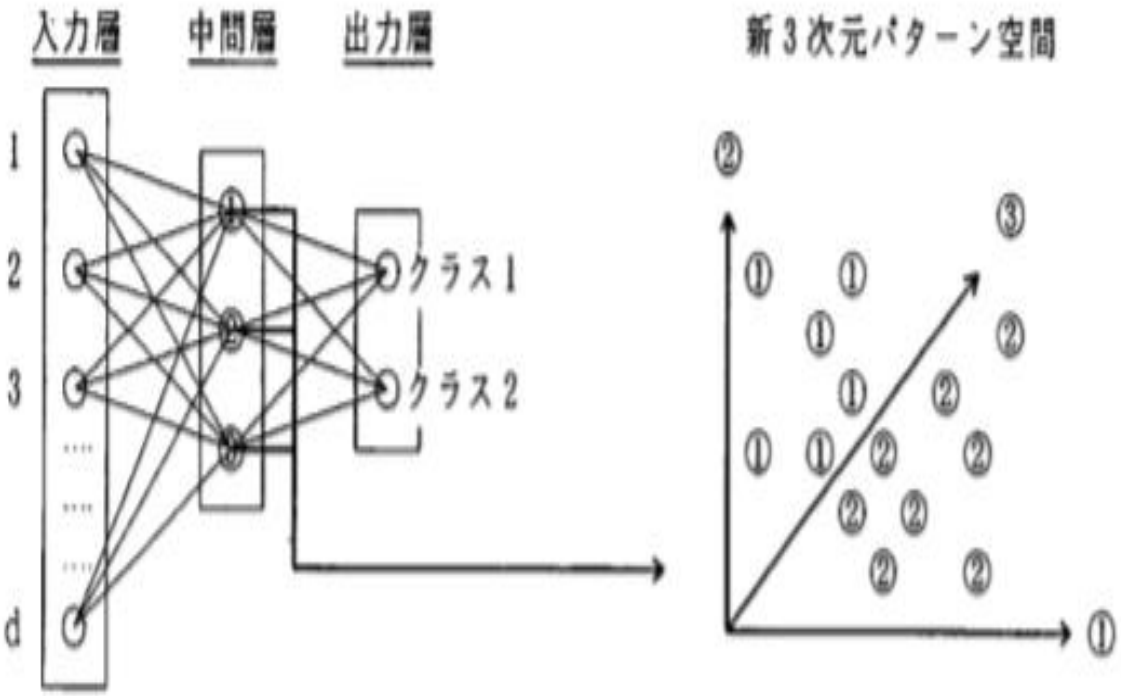


図５－７．バックプロパゲーションによる写像。 d次元⇨３次元
（新たな次元は各パターンがクラス１とクラス２とに分類されやすいように
３次元空間上に分布していることに注意）

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

アンサンブル学習法 Ensemble learning method

AdaBoostやランダムフォレスト等で採用されている解析手法
Analysis methods used in AdaBoost and Random Forest

特徴:
弱分類機を条件を変えつつ多数用い、最終的な分類結果は
個々の分類結果データを統合して最終結果とする。
A number of weak classifiers are used while changing the conditions, and the final classification result is obtained by integrating individual classification result data

従来の分類手法を組み合わせて使う「メタ解析手法」である。
＊「KY法」もメタ解析手法となる
This is a “meta-analysis method” that uses a combination of conventional classification methods.
"KY method" is also a meta-analysis method

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

◇ニューラルネットワーク Neural network

＊ニューラルネットワークは単純なネットワーク構造を利用したパーセプトロンを基本として発展
＊パーセプトロンは簡単な二分類問題も解決できないということで、衰退
＊パーセプトロンの限界を打破したアプローチとしてニューラルネットワークが提案

* Neural networks are developed based on perceptrons that use a simple network structure.
* Perceptron declines because it cannot solve simple two-classification problems
* A neural network is proposed as an approach that breaks the limits of perceptron.

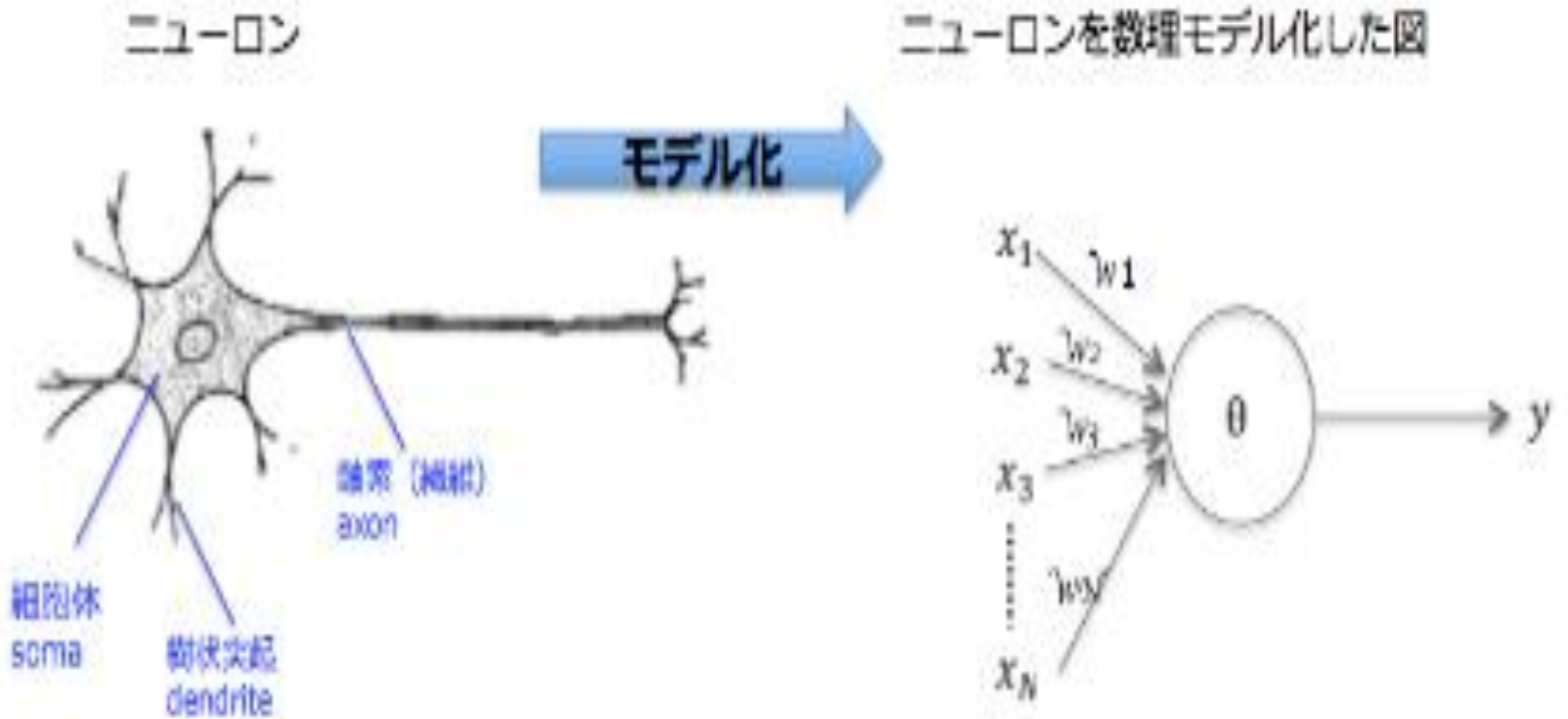
パーセプトロン：線形分類機 Perceptron: linear classifier

ニューラルネットワーク：非線形分類機
Neural network: nonlinear classifier

＊パーセプトロンは脳の機能を模したアプローチとして開発され、最終目標はAI(人工知能）への展開であった
＊ニューラルネットワークはその改良系で、ネットワーク構造が多層構造となった
＊最近の深層学習はニューラルネットワークのネットワーク構造をさらに複雑にした

* Perceptron was developed as an approach that mimics the function of the brain, and the final goal was to develop AI (artificial intelligence)
* Neural network is an improved version of the network structure.
* Recent deep learning has further complicated the network structure of neural networks.

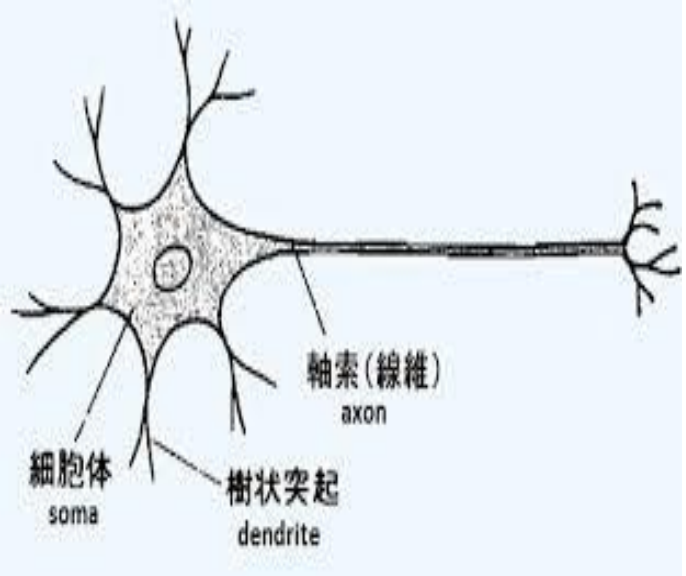
ロパーセプトロン ： perceptron



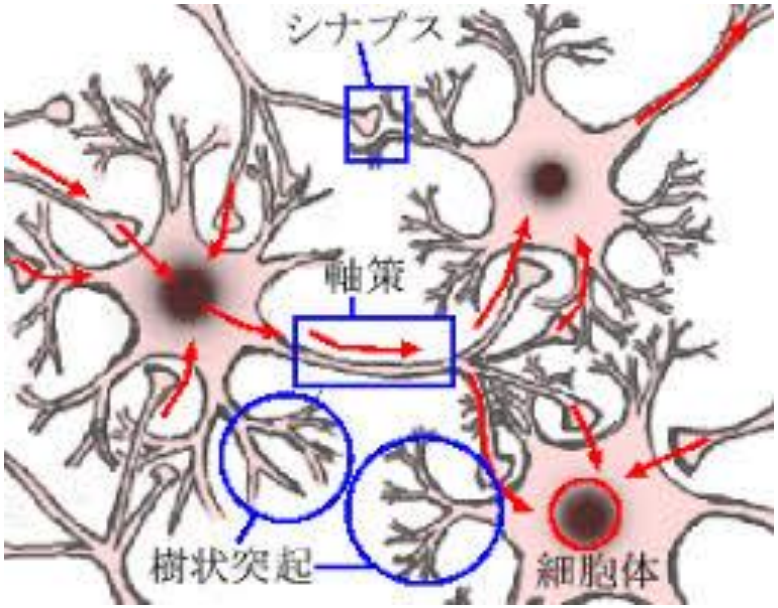
https://udemy.benesse.co.jp/ai/neural-network.html

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

単ニューロンモデル
Single neuron model



ネットワークニューロンモデル
Network neuron model



http://www.tamagawa.ac.jp/teachers/aihara/kouzou.html

http://www.sys.ci.ritsumei.ac.jp/project/theory/nn/nn.html

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

パーセプトロン
perceptron

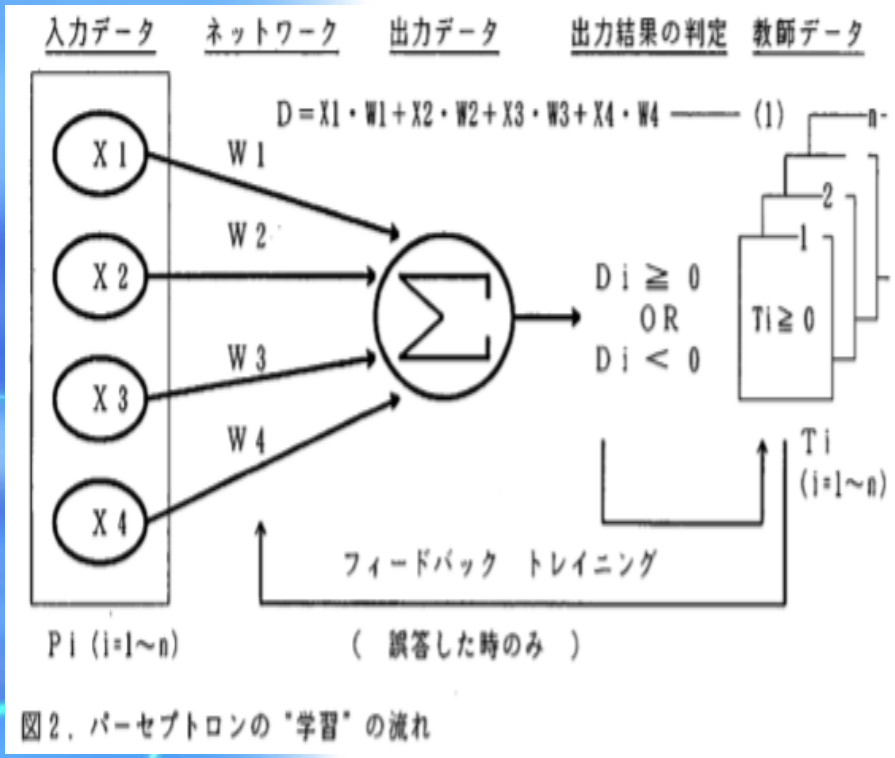
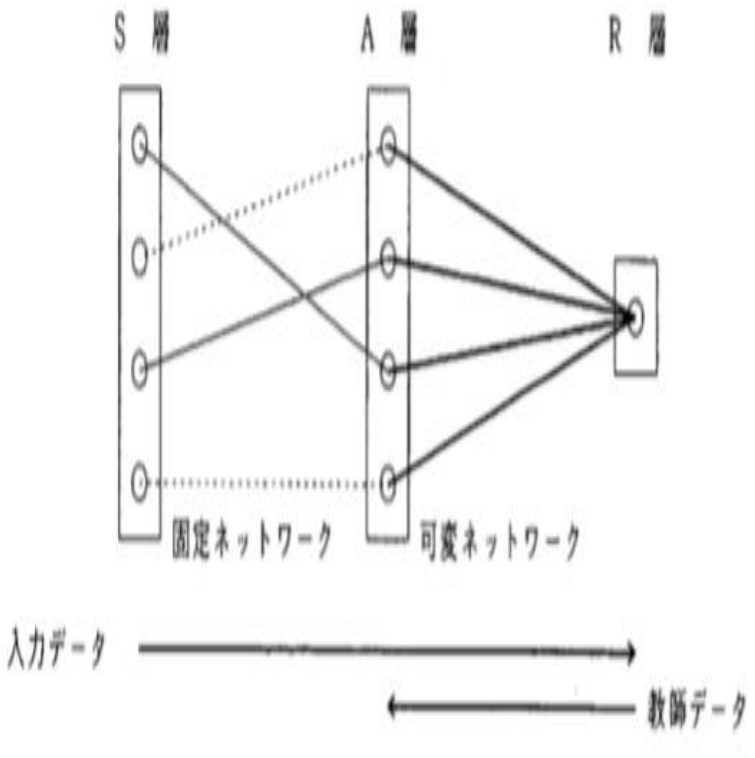


図2．パーセプトロンの"学習"の流れ

単ニューロンモデル
Single neuron model

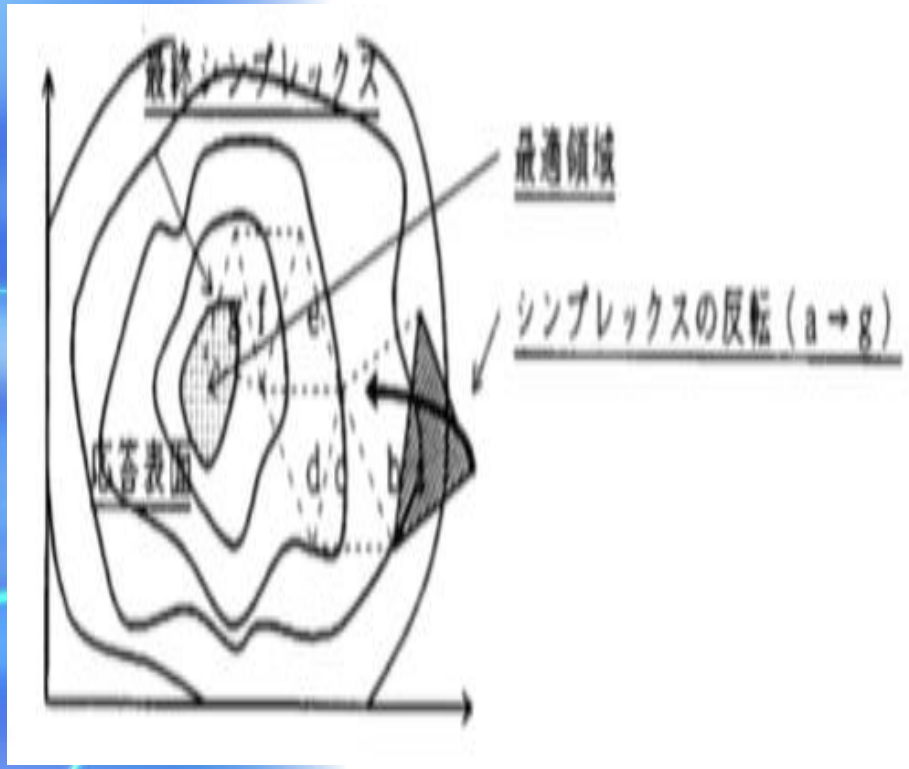
ニューラルネットワーク
neural network



ネットワークニューロンモデル
Network neuron model

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

◇最適化手法：シンプレックス法
Optimization method: Simplex method



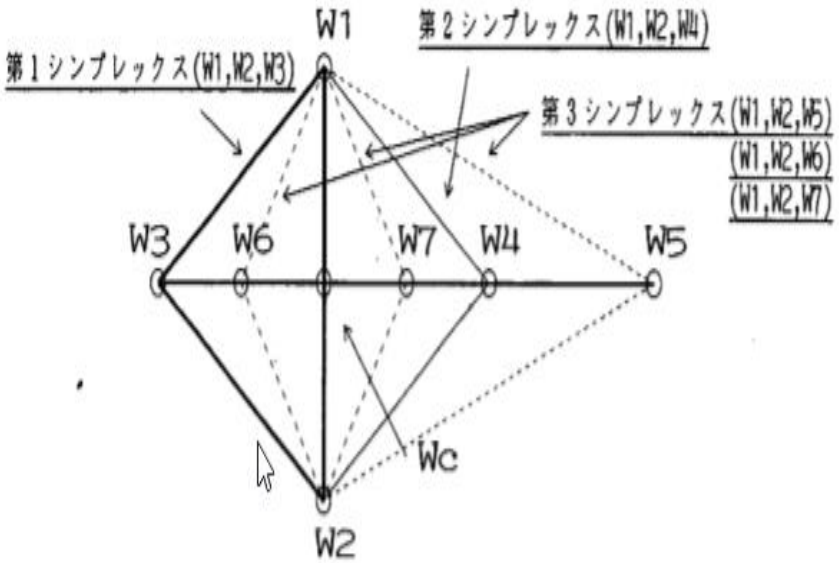


図6－2　シンプレックス反転に関するルール

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

◇最適化手法：遺伝的アルゴリズム
Optimization method: Genetic algorithm

遺伝子の分裂／増殖／突然変異等の動きをシミュレーションするアプローチ
Approach for simulating gene division / growth / mutation

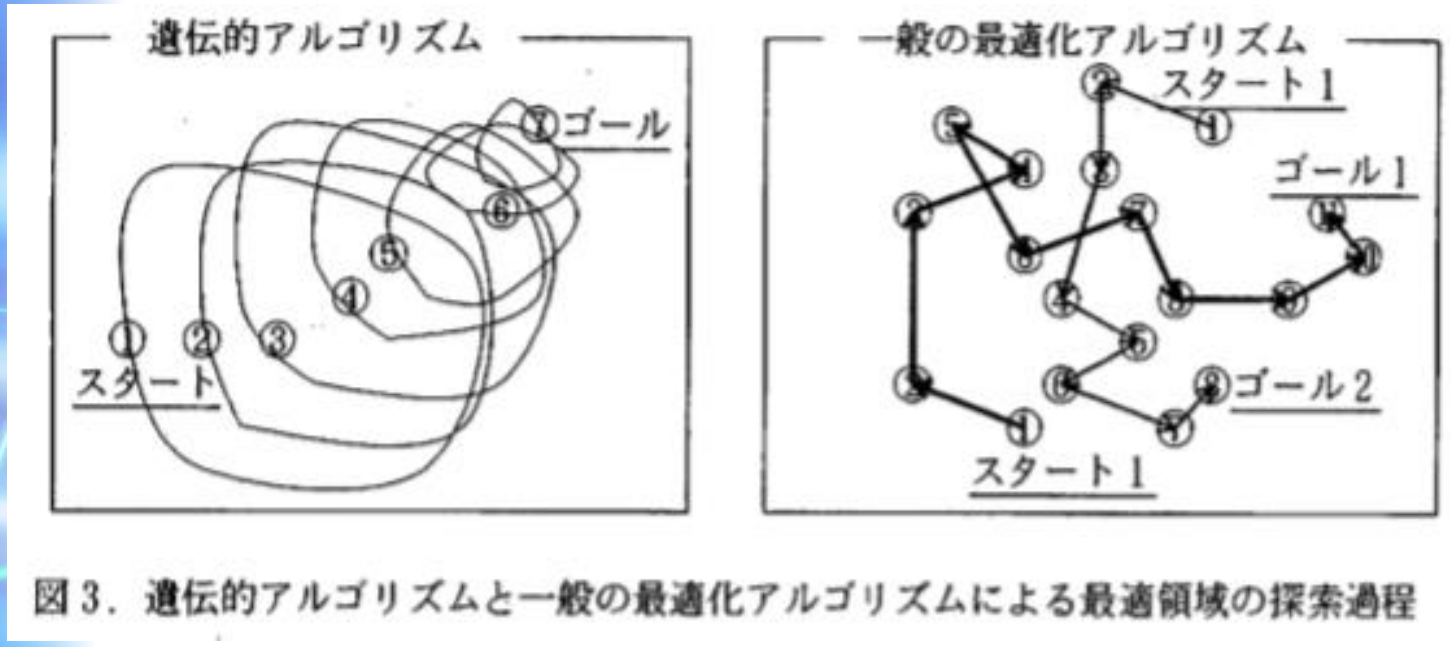


図3．遺伝的アルゴリズムと一般の最適化アルゴリズムによる最適領域の探索過程

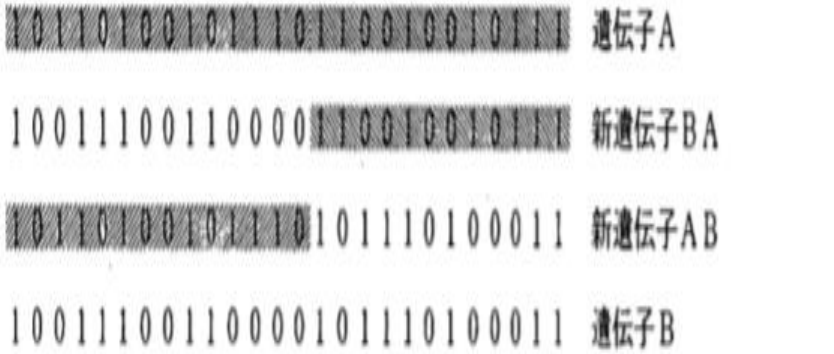
1．増殖（MULTIPLICATION）
2．交叉（CROSSING-OVER）
3．突然変異（MUTATION）
4．淘汰／選択（SELECTION）

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

◇最適化手法：遺伝的アルゴリズム
Optimization method: Genetic algorithm

1. 増殖 (MULTIPLICATION)

2. 交叉 (CROSSING-OVER)
情報の取替により、情報の変換を果たすものである。



異なる遺伝子パターンをもつ2本の遺伝子がある時、これら2本の遺伝子を交叉させる。 この結果、新たに2本の遺伝子が誕生するがこれらの遺伝子は親となるAおよびBの遺伝子の形質を交叉させた点を中心としてそれぞれあわせ持つことがわかる。

3. 突然変異 (MUTATION)
この突然変異では遺伝子の持つ情報がある部位で変化して対立遺伝子となる。

4. 淘汰／選択 (SELECTION)
様々な変換パターンにより構築された遺伝子が淘汰により悪い形質（問題解決に取って望ましくない）を持つものが取り除かれてゆくことである。 生物学的にはまわりの環境に適合した形質を持つ固体だけが生き残り（選択され）、次世代へと情報（形質）を伝えてゆくことを意味する。

◇個別手法の特徴と適用内容
Characteristics and application contents of individual methods

◇最適化手法：ニューラルネットワーク
Optimization method: Neural network

閉じたネットワーク構造を有するニューラルネットワーク
Neural network with closed network structure
ボルツマンマシンおよびホプフィールドネット
Boltzmann machine and Hopfield net

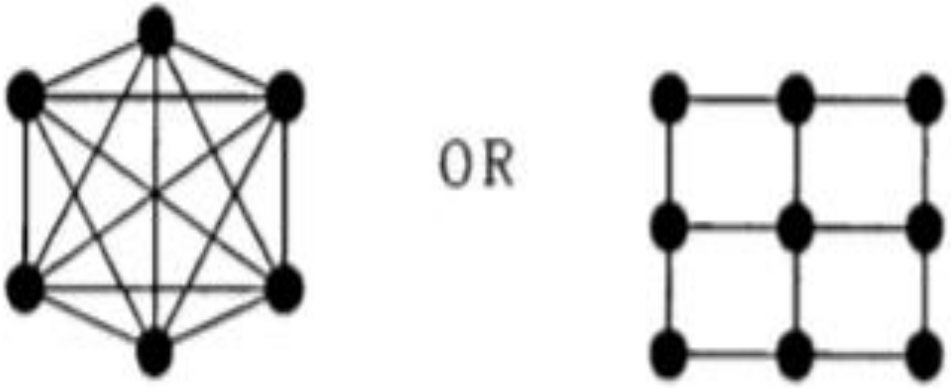


図1．ボルツマンマシン及びホップフィールドネットのネットワーク構造

◇KY法について
K-step Yard sampling methods

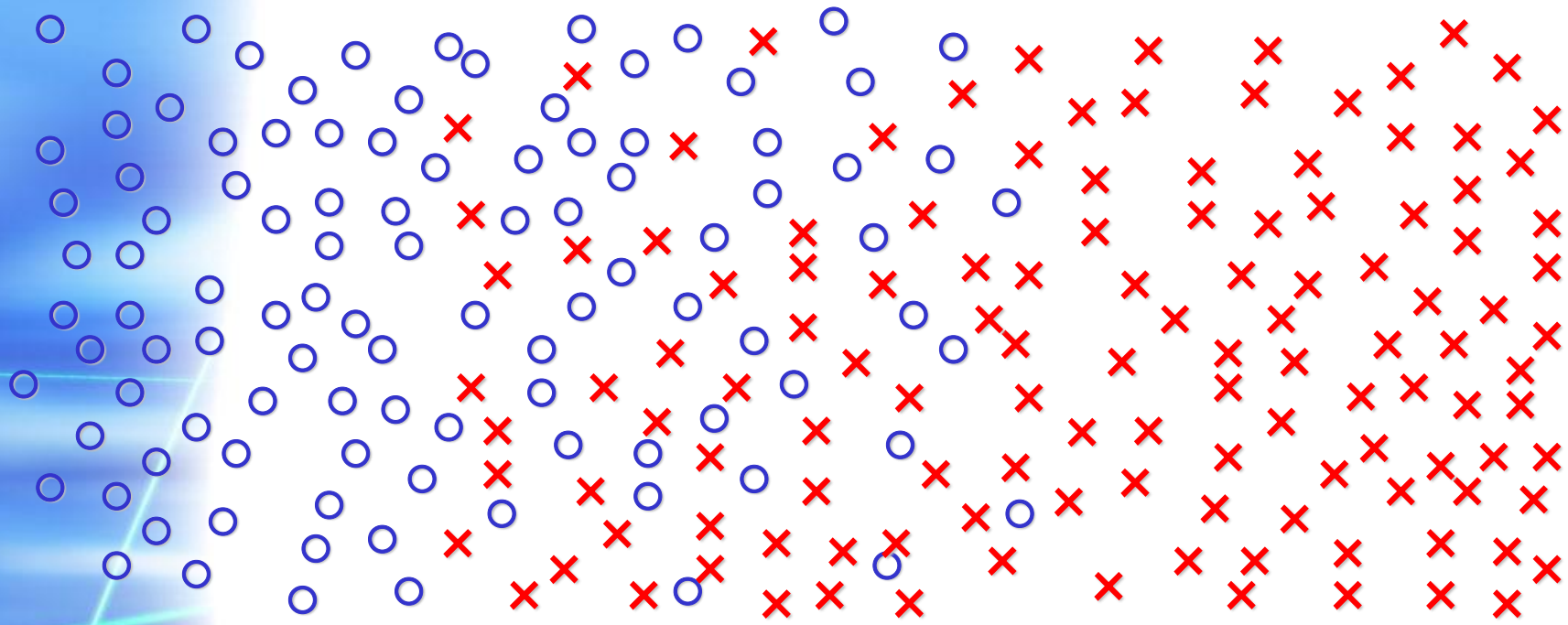
KY法(K-step Yard sampling methods)の特徴:以下の二大特徴を持つとKY法となる
Features of KY method (K-step Yard sampling methods): KY method with the following two major features

①多段階解析手法:一回の解析でなく、サンプル群を変えて何回も解析する
Multi-stage analysis method: Analyze many times by changing the sample group, not a single analysis
②メタ解析手法;KY法の各ステップで使われるデータ解析手法は従来法である
Meta-analysis method; the data analysis method used in each step of the KY method is the conventional method

①の特徴により、各段階においてサンプル群のリサンプリングが実施される。
Sample group resampling is implemented at each stage due to the characteristics of ①.
繰り返しデータ解析が実施される点でアンサンブル法に似ているが、
Similar to the ensemble method in that repeated data analysis is performed
・アンサンブル法、適用対象となるサンプル母集団は変化しない
Ensemble method, sample population to be applied does not change
・KY法では、サンプル母集団が変化し、これに対してデータ解析が実施される
In the KY method, the sample population changes, and data analysis is performed against this

②メタ解析手法という点でアンサンブル法と同じであるが、以下の点で大きく異なる
Same as ensemble method in terms of meta-analysis method, but greatly different in the following points
・KY法ではステップ単位で適用されるデータ解析手法を変えることが出来る
The KY method can change the data analysis method applied on a step-by-step basis
・アンサンブル法では、基本的に適用されるデータ解析手法は同じ手法を用いる
In the ensemble method, basically the same data analysis method is used.

◆ 一般的なサンプル空間：General sample space



○：Positive

×：Negative

100%分類の実現は極めて困難

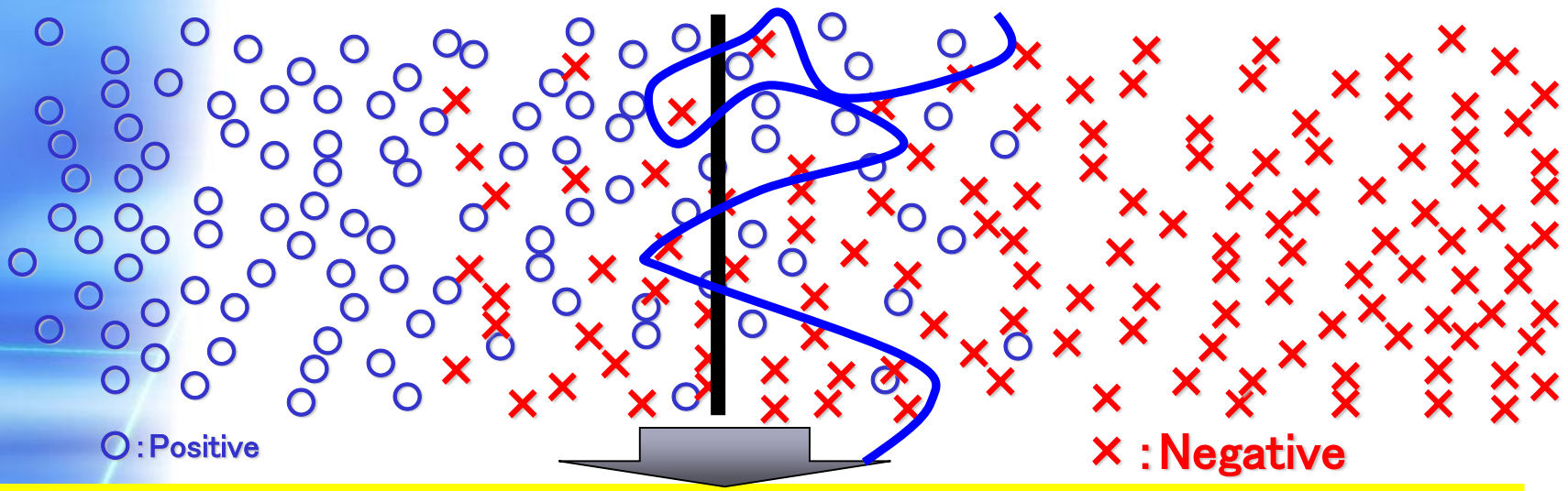
Realization of 100% classification is almost impossible

◆パターン分布とパターン分類：識別不可能な場合

Pattern distribution and pattern classification: When identification is impossible

100%分類不可能な識別線（判別関数）

Discrimination line that cannot be classified 100% (discriminant function)



新たなサンプル空間を創出する、非線形化する、
次元の拡張等行ったとしても100%分類の可能性は殆ど無い

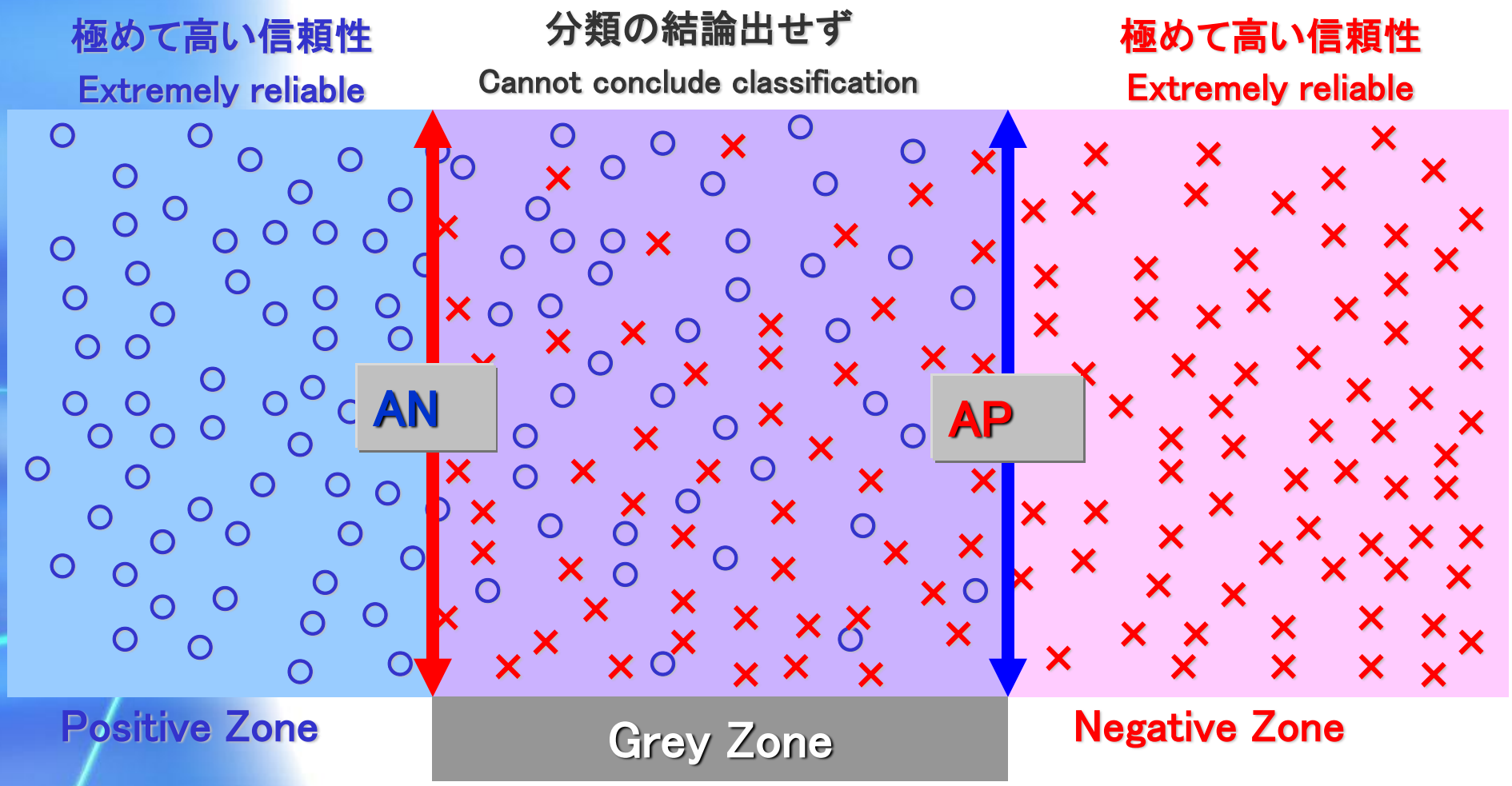
There is almost no possibility of 100% classification even if a new sample space is created, delinearized, or dimension extended.

100%分類の実現は極めて困難

Realization of 100% classification is extremely difficult

AP／AN判別関数の組み合わせによる分類特性

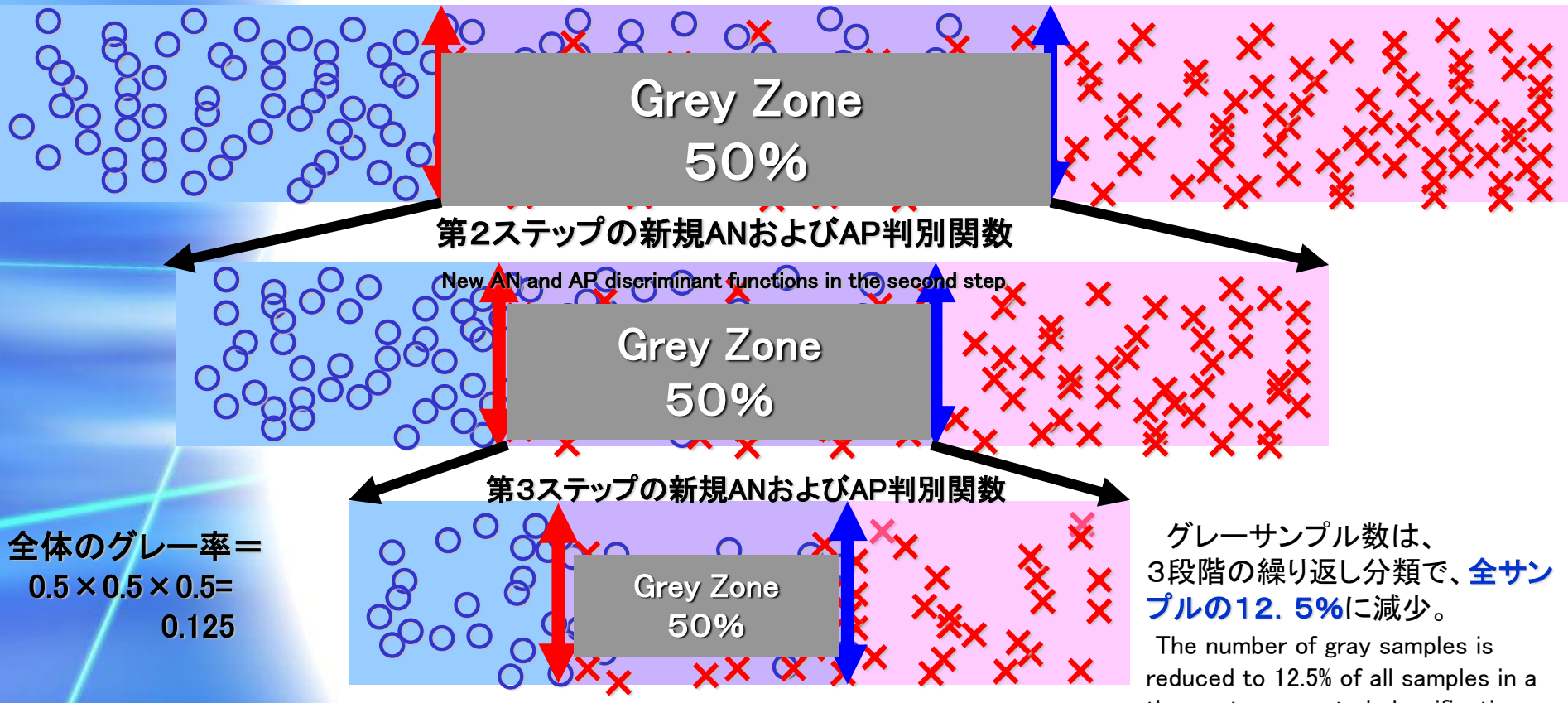
Classification characteristics by combination of AP / AN discriminant functions



グレーゾーンを取り出して再分類を繰り返すアプローチ

An approach that takes out gray zones and repeats reclassification

“K-step Yard sampling（KY）法”



全体のグレー率＝
0.5×0.5×0.5=
0.125

グレーサンプル数は、
3段階の繰り返し分類で、**全サンプルの12．5%**に減少。

The number of gray samples is reduced to 12.5% of all samples in a three-step repeated classification.

□ KY(K-step Yard sampling)法について

Challenge for classification and prediction

K-step Yard sampling methods

KY-methods

↓

The most powerful and advanced data analysis method

↓

The most difficult classification problem
6,965 sample of Ames test samples were,
Classified perfectly

Application test of "K-step Yard sampling"

☐Samples

1. Ames test data
2. Sample population
   total :6,965
   Mutagen; 2,932
   Non-mutagen; 4,033

☐Result of KY-method

1. Number of steps : 23 steps ; 22 (2 models) + 1 (1 model)
2. Classification ratio : 100 %

☐Used system

ADMEWORKS / ModelBuilder　V 3.0.22

☐Used parameters (Initial condition)

Number of generated parameters : 838
Number of parameters for step 1 : 98
Confidency index (Samples(6965) / Parameters(98)) : 71.1 > 4.0

Application test by normal and various D.A. methods

1. Linear discriminant analysis with linear least-squares method

Classification ratio : total; 73.50(6965), Mutagen;73.02(2932), Non mutagen;73.84(4033)
Number of mis-classified : (1846), (791) (1055)
Prediction ratio (L100 out) 72.58% deviance(0.92%)
(L500 out) 73.32% deviance(0.18%)

2. SVM (Support Vector Machine with Kernel)

Classification ratio : total; 90.87(6965), Mutagen;86.83(2932) Non mutagen; 93.80(4033)
Number of mis-classified : (636), (386) (250)
Prediction ratio (L500 out) 80.99% deviance(9.88%)

3. AdaBoost

Classification ratio : total; 77.24(6965), Mutagen;66.13(2932) Non-mutagen; 85.32(4033)
Number of mis-classified : (1585) (993) (592)
Prediction ratio (L500 out) 75.16% deviance(2.08%)

Classification result by the AdaBoost

Classification ratio of total 6,965 compounds is 77.24%

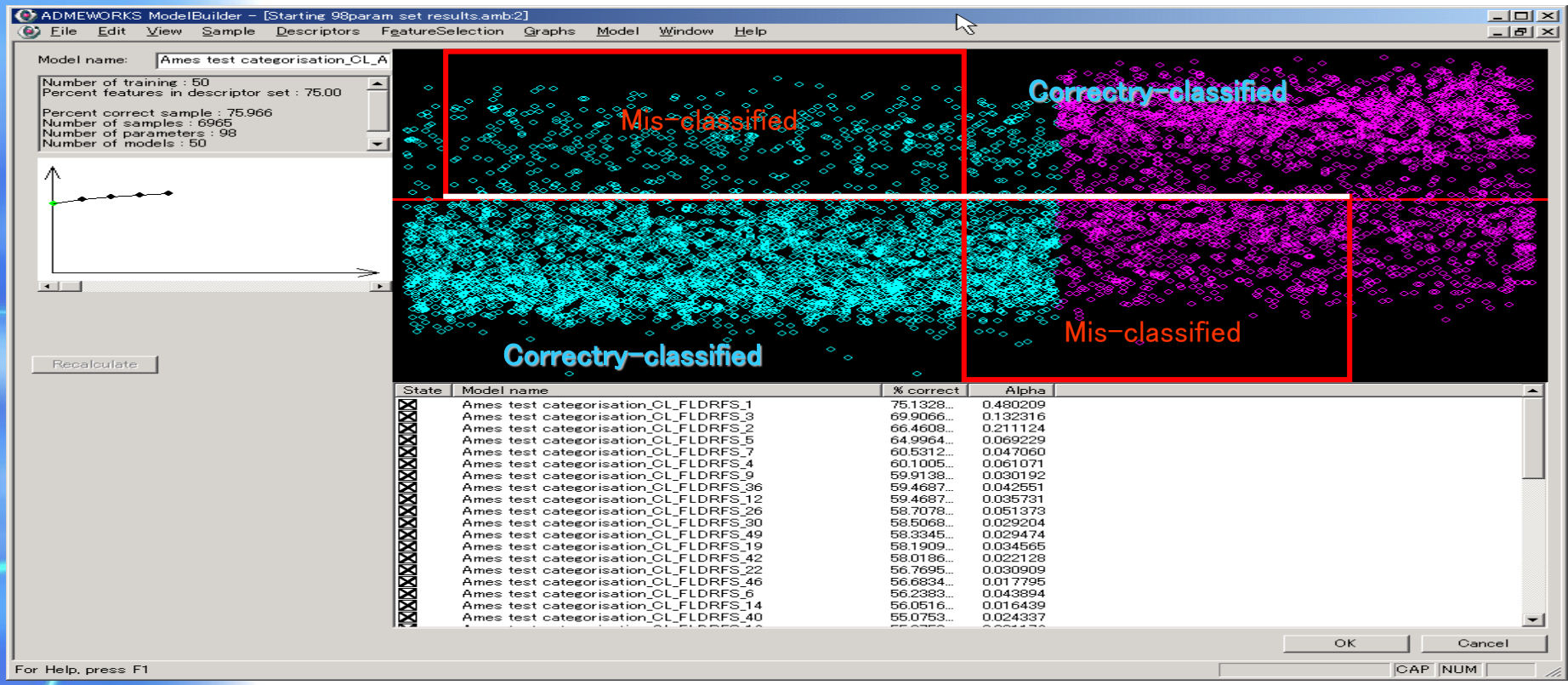
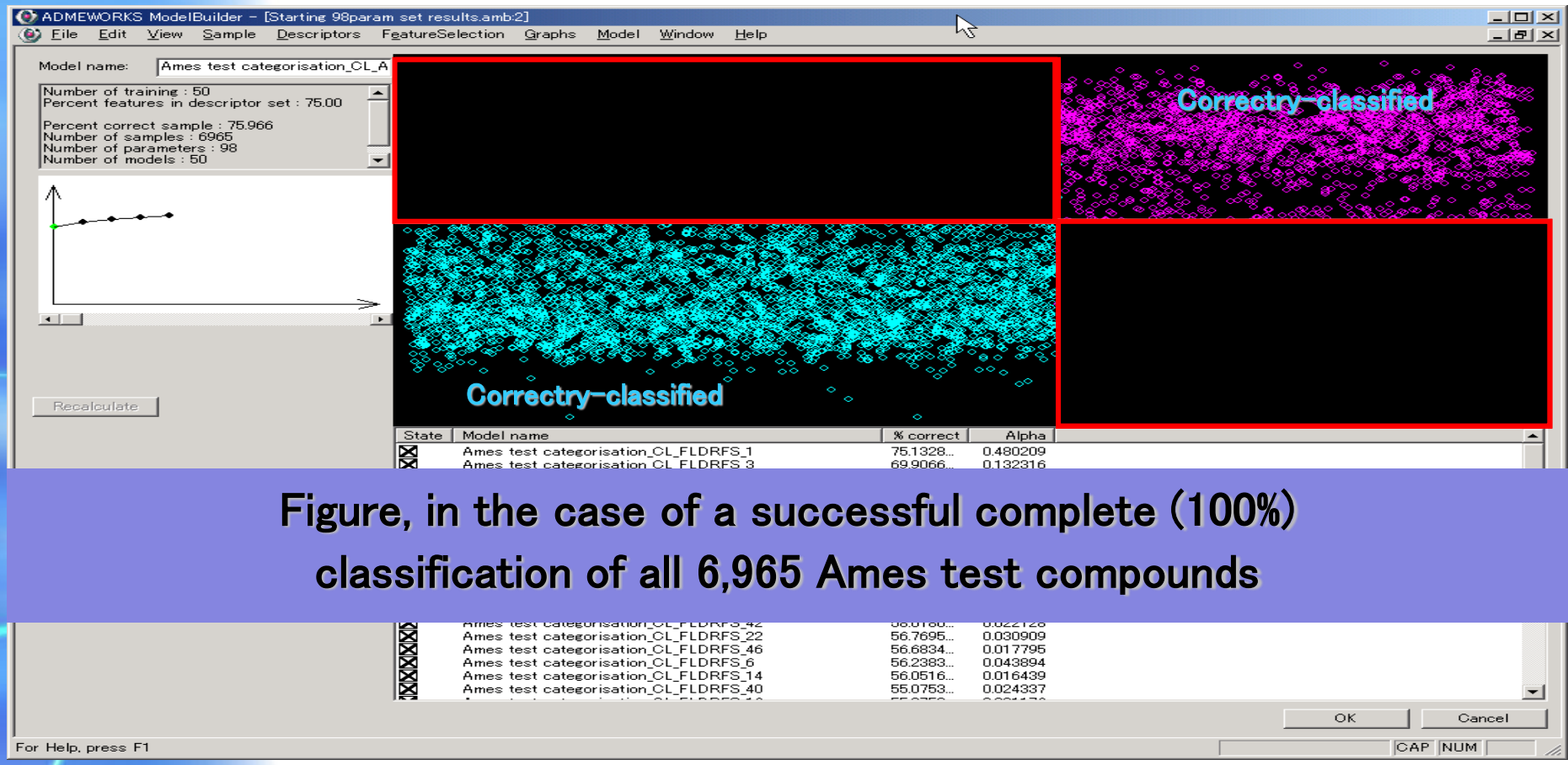


Figure of the ADMEWORKS

Classification result by the AdaBoost



Figure, in the case of a successful complete (100%) classification of all 6,965 Ames test compounds

Figure of the ADMEWORKS

“K-step Yard sampling (KY) Method”

Total steps : 23 steps (2 models) + 1 step (1 model)

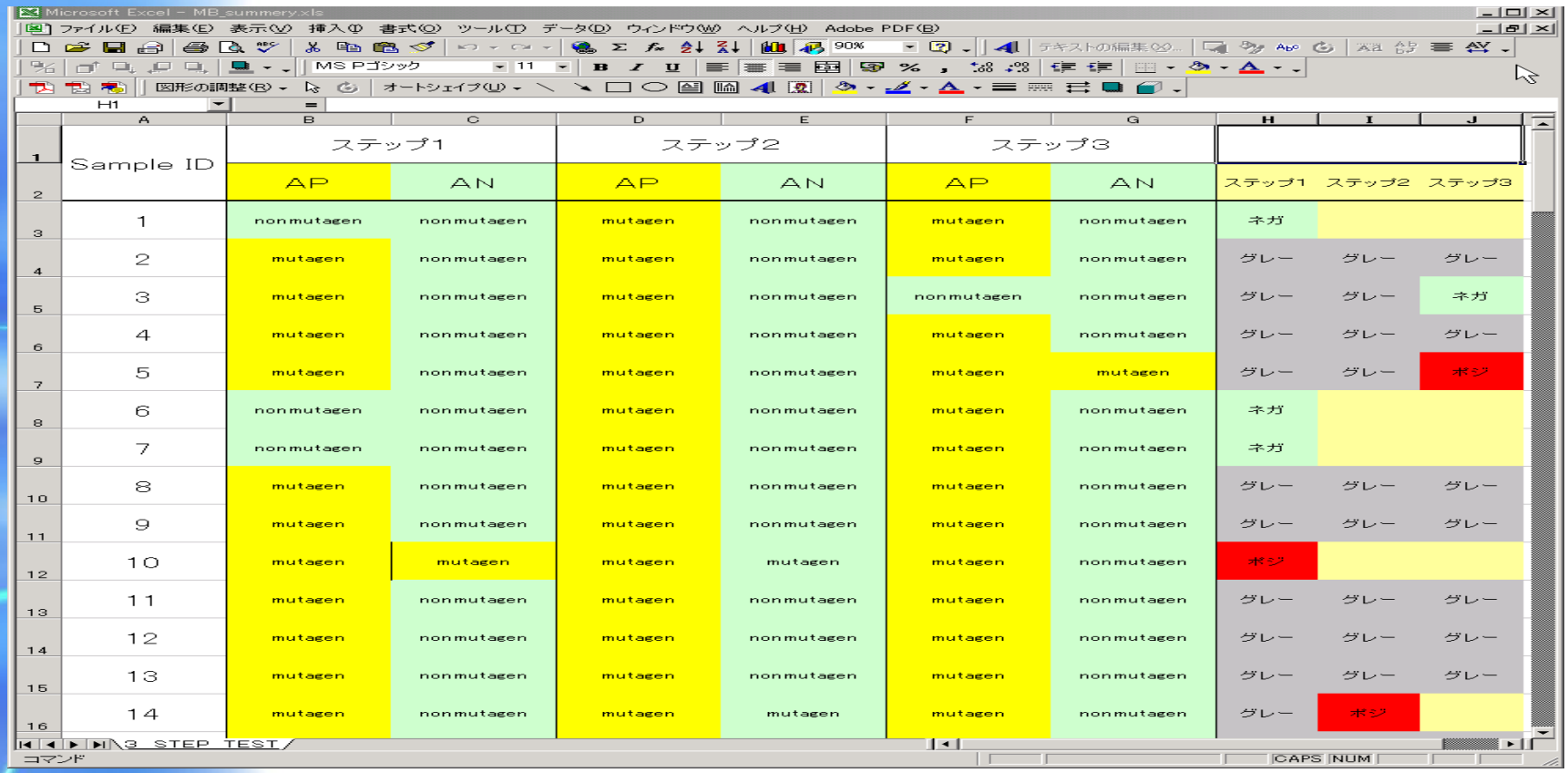
ステップID（KY法）	Starting samples(Total) Final samples Determined samples(Total)	Mutagen (Initial) Mutagen (Final) Determined samples(Mut.)	Non-mutagen (Initial) Non-mutagen (Final) Determined samples(Non-mut.)	Grey sample (Initial) Grey sample (Final) Grey ratio(%) (Grey/Total)
1	6965	2932	4033	0
	5864	2413	3451	5864
	1101	519	582	84.19
2	5864	2413	3451	5864
	5108	2142	2966	5108
	756	271	485	87.11
3	5108	2142	2966	5108
	4486	1919	2567	4486
	622	223	399	87.82
4	4486	1919	2567	4486
	4133	1779	2354	4133
	353	140	213	92.13
5	4133	1779	2354	4133
	3794	1651	2143	3794
	339	128	211	91.8
6	3794	1651	2143	3794
	3462	1485	1977	3462
	332	166	166	91.25
7	3462	1485	1977	3462
	3090	1345	1745	3090
	372	140	232	89.25
8	3090	1345	1745	3090
	2826	1220	1606	2826
	264	125	139	91.46
9	2826	1220	1606	2826
	2592	1139	1453	2592
	234	81	153	90.63
10	2592	1139	1453	2592
	2384	1047	1337	2384
	208	92	116	91.98

“K-step Yard sampling (KY) Method”

12	2095	931	1164	2095
	1848	829	1019	1848
	247	102	145	88.21
13	1848	829	1019	1848
	1607	733	874	1607
	241	96	145	86.96
14	1607	733	874	1607
	1380	623	757	1380
	227	110	117	85.87
15	1380	623	757	1380
	1028	466	562	1028
	352	157	195	74.49
16	1028	466	562	1028
	787	358	429	787
	241	108	133	76.56
17	787	358	429	787
	529	234	295	529
	258	124	134	67.22
18	529	234	295	529
	392	201	191	392
	137	33	104	74.1
19	392	201	191	392
	279	141	138	279
	113	60	53	71.17
20	279	141	138	279
	184	105	79	184
	95	36	59	65.95
21	184	105	79	184
	112	66	46	112
	72	39	33	60.87
22	112	66	46	112
	66	39	27	66
	46	27	19	58.93
	66	39	27	66
23(1 model)	0	0	0	0
	66	39	27	0

“K-step Yard sampling (KY) Method”

Classification results by 3 steps



Sample ID	ステップ1		ステップ2		ステップ3				
	AP	AN	AP	AN	AP	AN	ステップ1	ステップ2	ステップ3
1	nonmutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	ネガ		
2	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
3	mutagen	nonmutagen	mutagen	nonmutagen	nonmutagen	nonmutagen	グレー	グレー	ネガ
4	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
5	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	mutagen	グレー	グレー	ポジ
6	nonmutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	ネガ		
7	nonmutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	ネガ		
8	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
9	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
10	mutagen	mutagen	mutagen	mutagen	mutagen	nonmutagen	ポジ		
11	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
12	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
13	mutagen	nonmutagen	mutagen	nonmutagen	mutagen	nonmutagen	グレー	グレー	グレー
14	mutagen	nonmutagen	mutagen	mutagen	mutagen	nonmutagen	グレー	ポジ	

Spatial features of "K-step Yard sampling"

☐Summary 1

■Features

1. KY-method is a meta-method
   Executed on various DA (linear and non-linear) methods.

2. Perfect classification is achieved on any condition

3. Three different types of KY-methods are available at this time

4. Applicable not only on Binary classification but Fitting methods

Spatial features of "K-step Yard sampling"

☐Summary 2

■Advantages

1. Sample number free approach

2. Sample distribution free approach

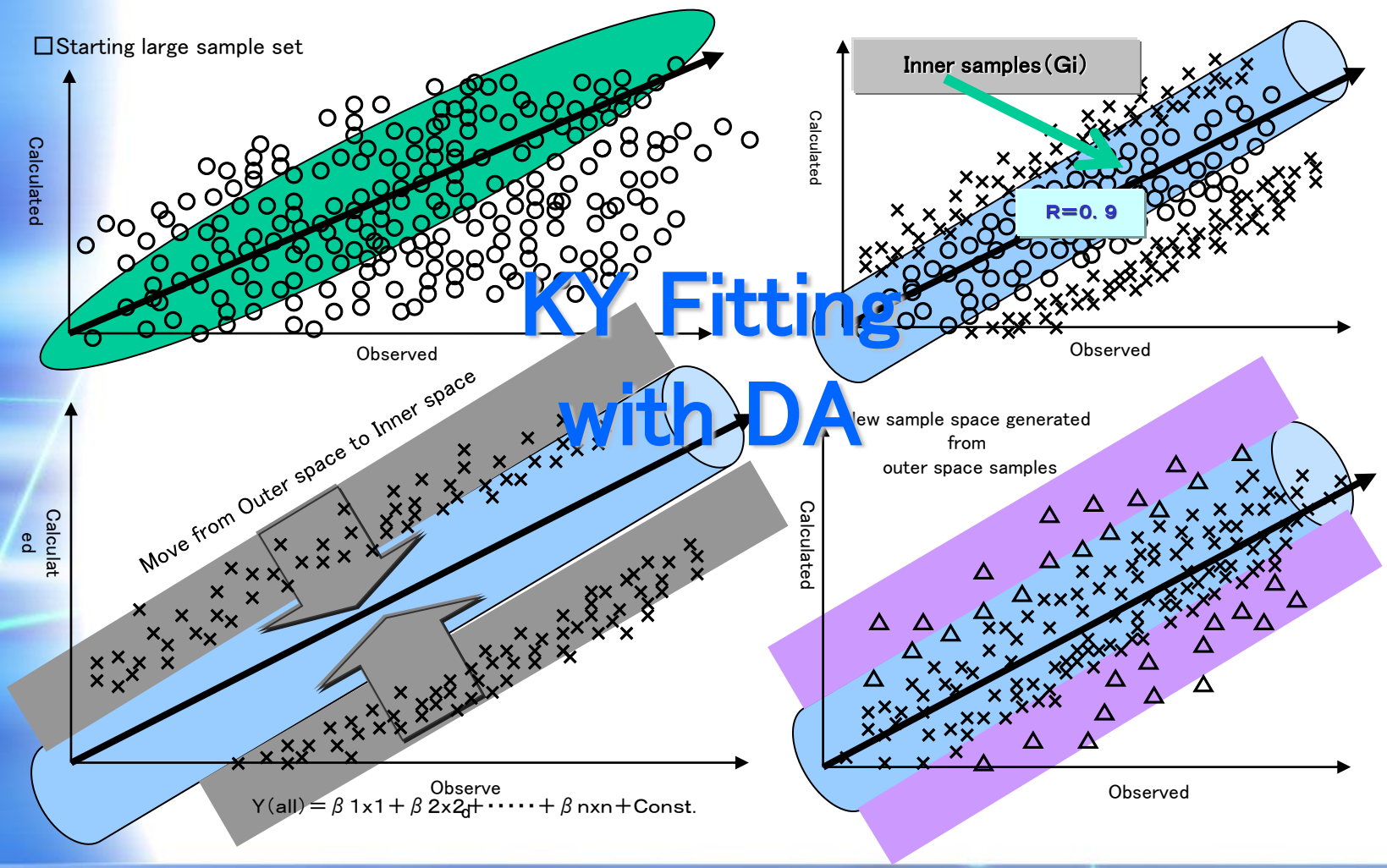
3. Perfect classification is achieved in any condition

4. Need not spatial data analysis softwares

■Disadvantages

Relatively complex operation to generate discriminant functions

◆KY method for fitting methods



◆KY method for fitting methods

Fish: 96 hours LC50、 Number of samples: 791、Log(1/LC50_Mm) (Max/Min) : 6.376 / −2.963

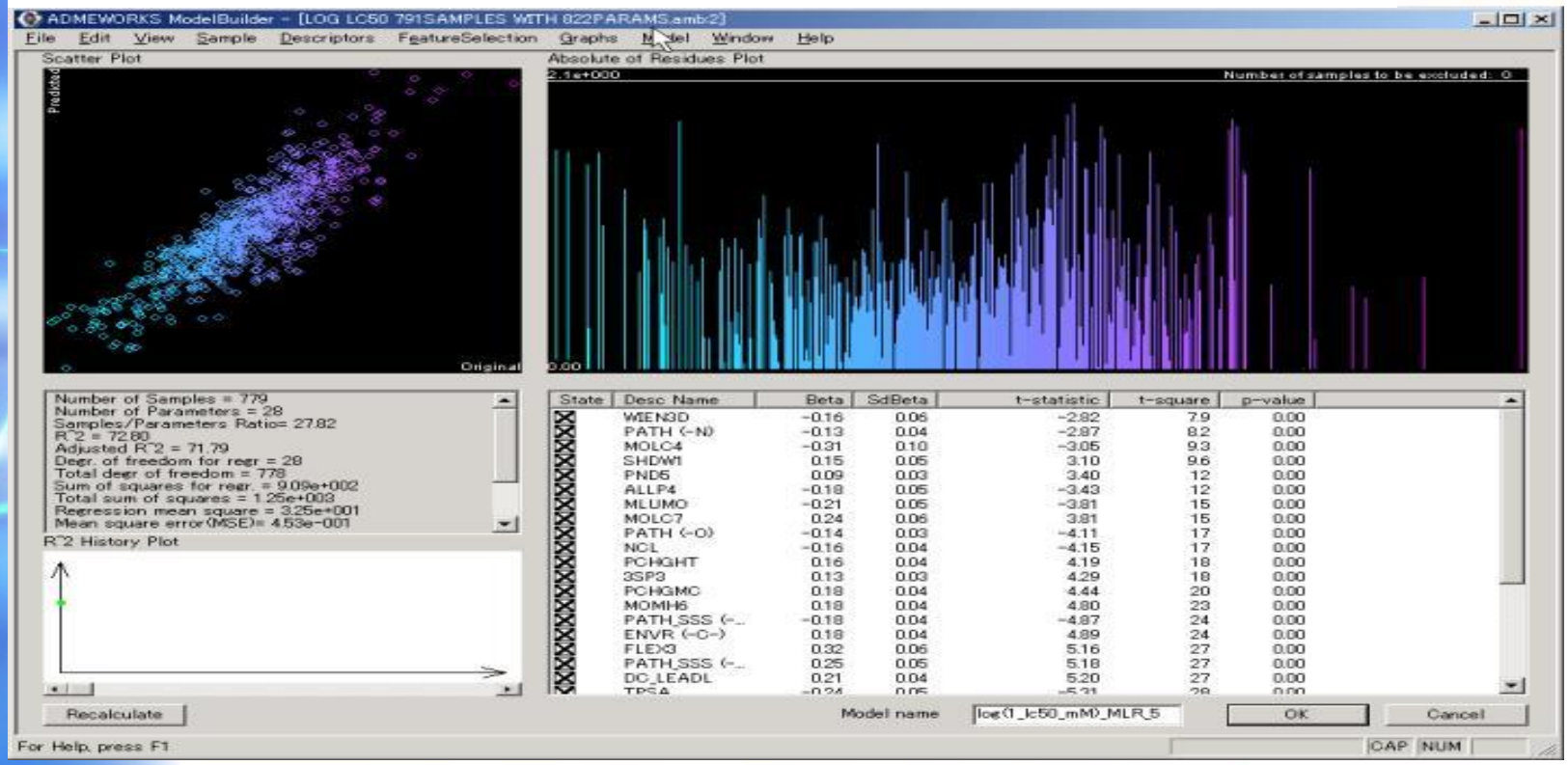
◇ Data analysis by ordinal linear

Step1 : Inner sample set

Number of samples : 779, Number of used parameters : 28, Confidance ratio : 27.8

R2 : 72.8, R : 85.3, F−value : 71.7, CV : 69.6

69. 6



◆フィッティングKY法実証実験

Fitting KY method demonstration experiment

◇フィッティングKY法による解析（ステージ1） Analysis by fitting KY method (Stage 1)

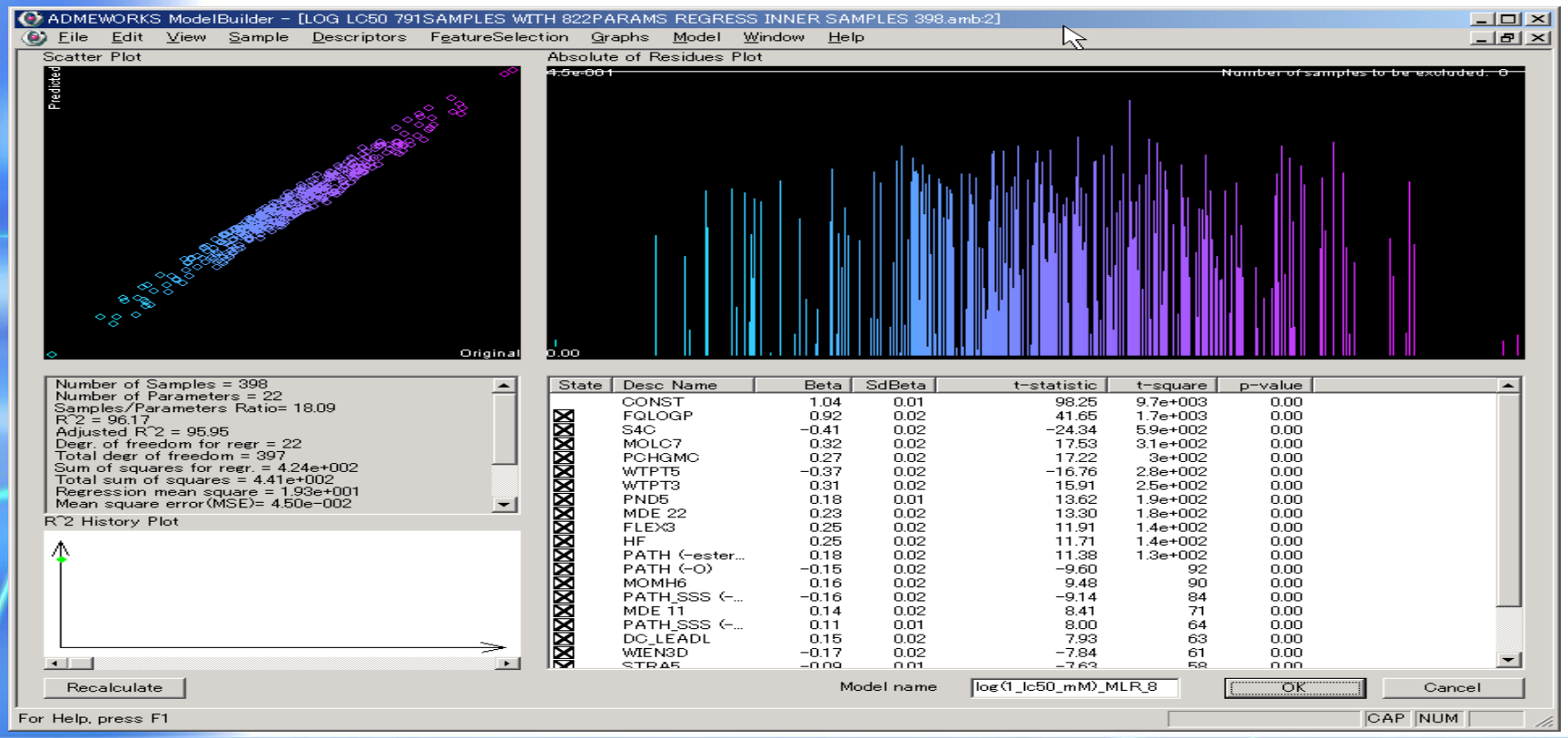
ステップ1：インナーサンプル Step 1: Inner sample

サンプル数：398、 パラメータ数：22、 信頼性指標：18．1

Number of samples: 398, number of parameters: 22, reliability index: 18.1

R2：96．2、 R：98．1、 F値：428、 クロスバリデーション：94．4

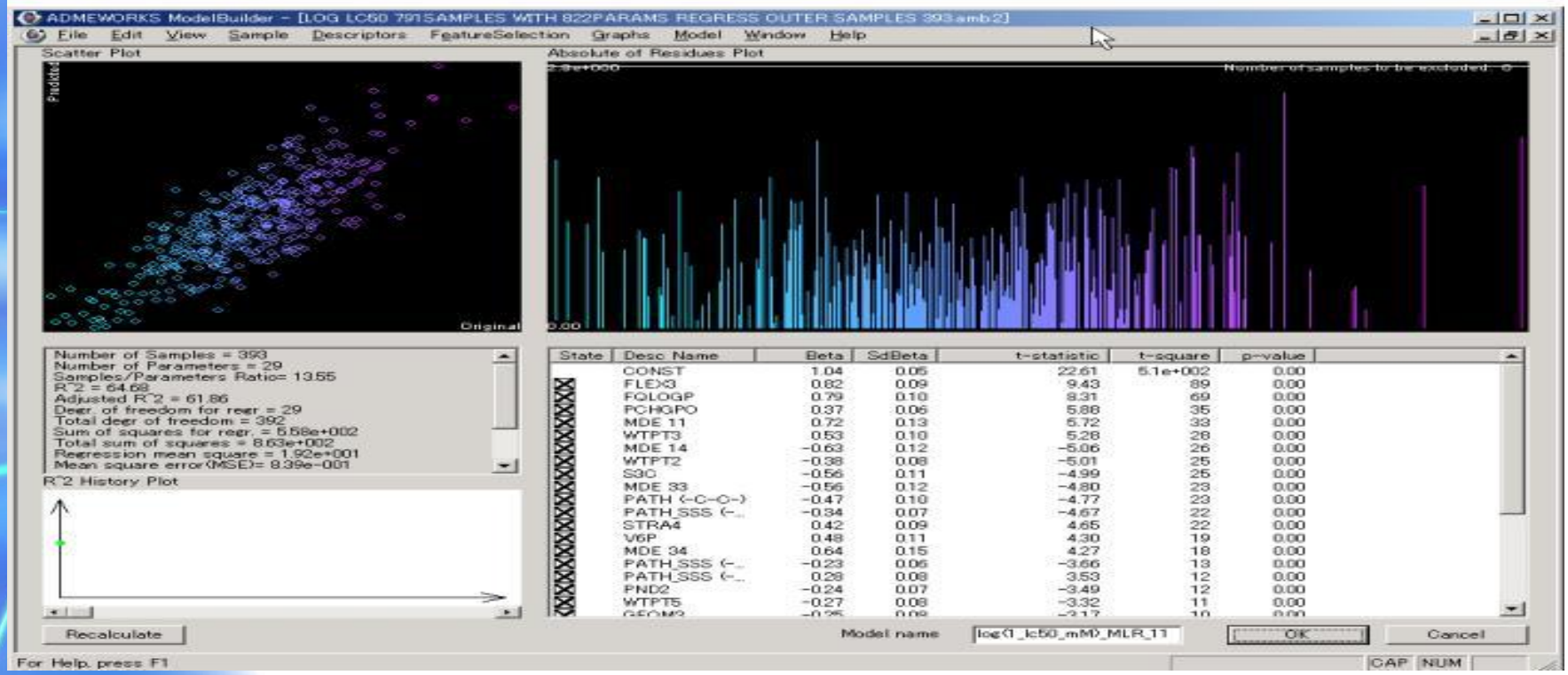
R2: 96.2, R: 98.1, F value: 428, Cross validation: 94.4



◆フィッティングKY法実証実験

◇フィッティングKY法による解析（ステージ1）

ステージ1： アウターサンプル
サンプル数：393、 パラメータ数：29、 信頼性指標：13．6
R2：64．7、 R：80．4、 F値：22.9、 クロスバリデーション：57．5



◇テーラーメードモデリング

＊予測モデルの精度は、モデル作成に用いたサンプルの質に大きく依存
＊予測モデルの精度向上にはサンプル母集団内のサンプルを目的解析に必要な情報を多く含むことが理想
＊解析目的と関係のある情報を含まないサンプルは極力避けるのが良い

＊化合物分野には「**似た化合物は似た活性／物性／毒性を有する**」という概念
＊予測対象化合物と似た化合物を選択して予測モデルを構築する

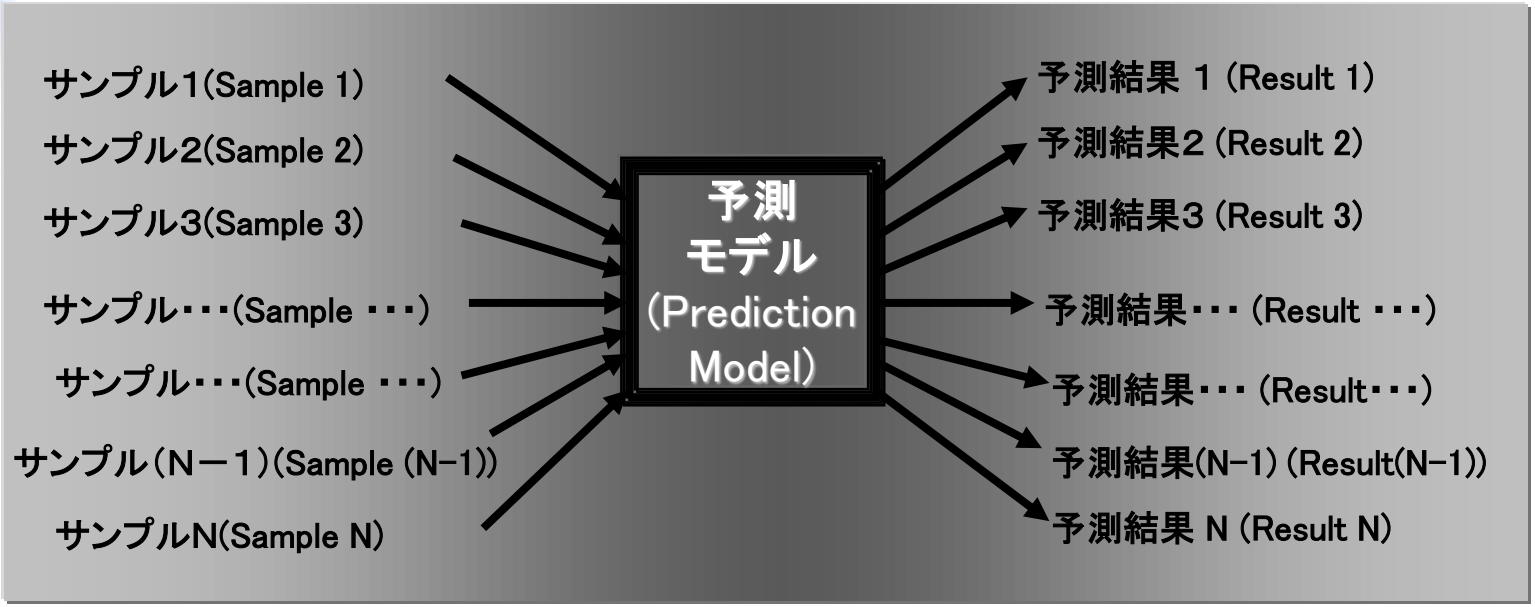
＊予測対象サンプル毎に、サンプル母集団を構成し、予測モデルを構築する

テーラーメードモデリング

United States Patent Application 20100145896 ; *Yuta June 10, 2010

従来手法による予測アプローチ
(Prediction approach by traditional method)

特徴：総てのサンプルを対象とした予測モデルの構築
Features:Generate a prediction model which can handle all samples



利点 (Merit):少ない数の予測モデル作成（Small number of prediction models are generated ）

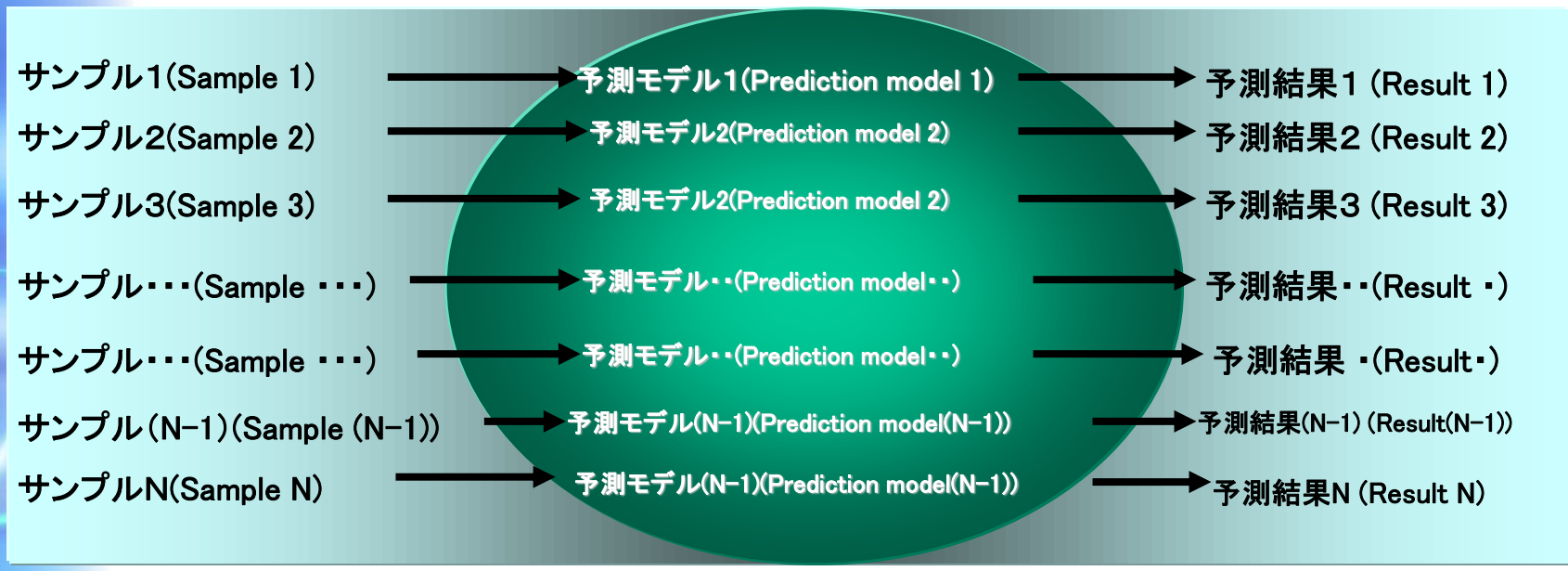
難点（Weakness）： 予測率の向上が困難である（Difficult to achieve high prediction ratio ）

「テーラーメード・モデリング」による予測アプローチ

(Prediction approach by “Tailor-Made Modeling”)

特徴:サンプル単位での予測モデルの構築

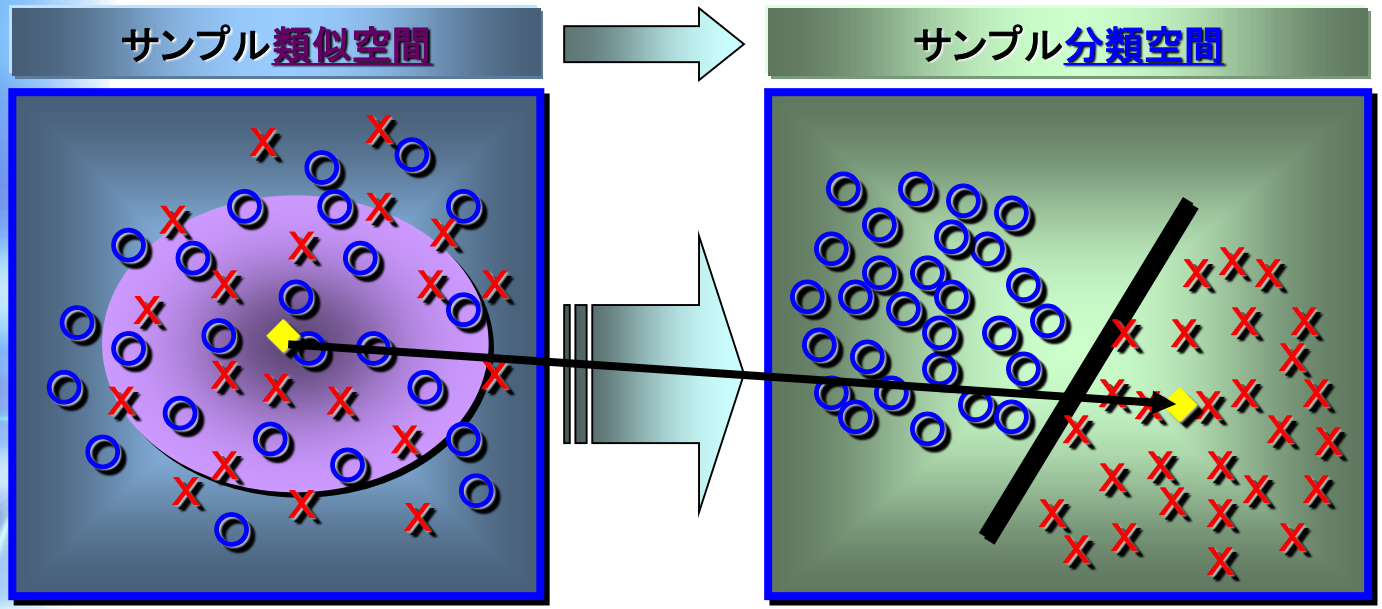
Features:Generate a prediction model which is designed for only 1 samples



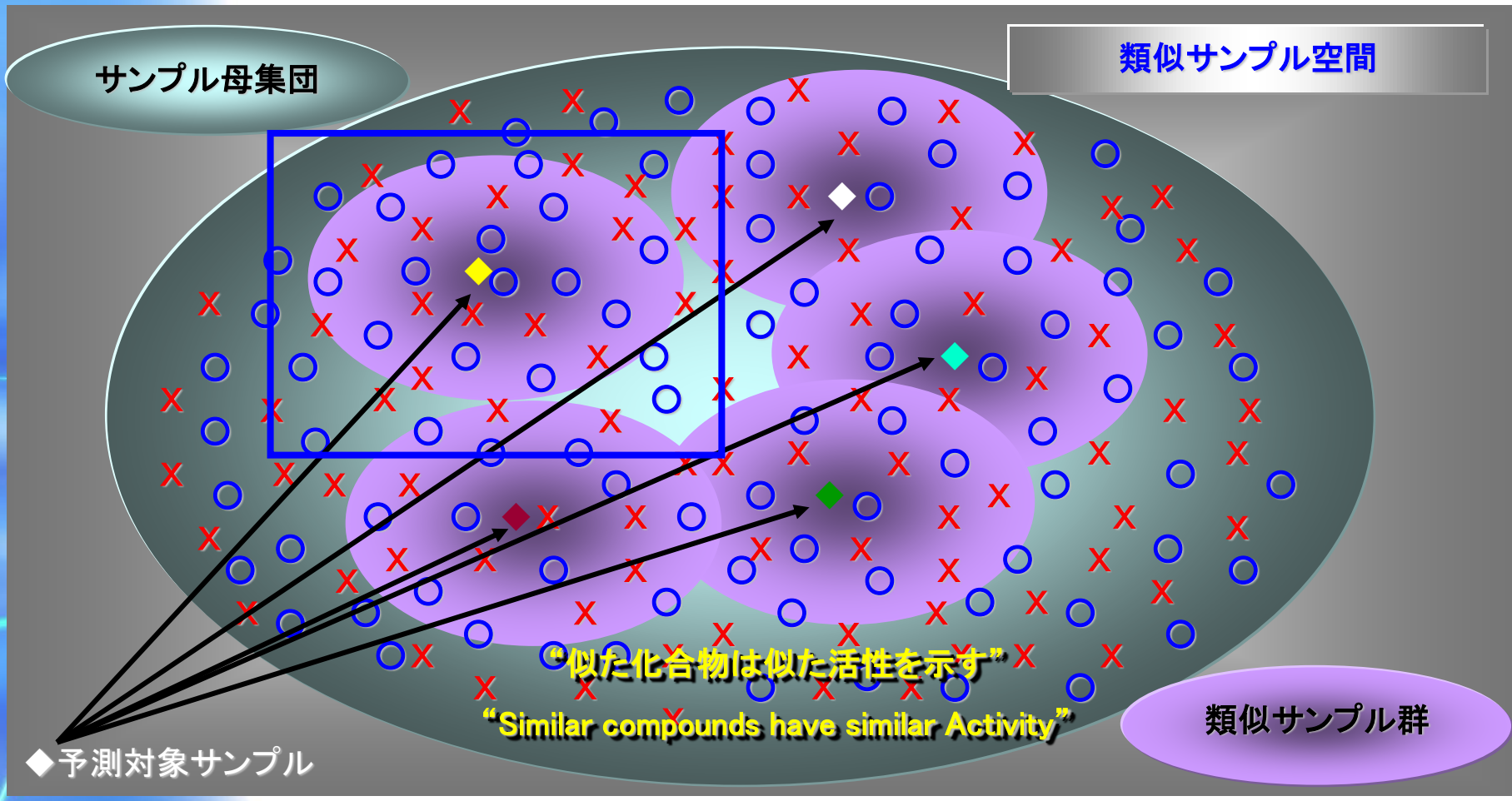
利点 (Merit)　　:　予測率が大幅に向上する (High prediction ratio will be achieved)

難点 (Weakness):　計算時間がかかる (Need large calculation time)

予測用サンプルの取り出しと、テーラーメード予測



サンプル母集団からの予測用サンプルの取り出し



◇ KY(K-step Yard sampling)法

■ **ニクラス分類KY法:Binary classification KY methods**
- ニモデルKY法; Two models BC KY method
- 一モデルKY法; One model BC KY method
- モデルフリーKY法; Model free BC KY method

■ **重回帰(フィッティング)KY法:Regression KY methods**
- 判別関数付き重回帰KY法;BC regression KY method
- 三ゾーン重回帰KY法;Three zone regression KY method
- モデルフリー重回帰KY法;Model free regression KY method

□ クラスタリングKY法; Clustering KY methods
□ 主成分KY法; Principal component KY methods

◇ 現在開発中
- リバーシKY法:ニクラス分類の簡略法
- ポピュレーションフリーKY法:ポピュレーション比率の悪い時に適用

◇ テーラーメードモデリング(Tailor made modeling)

■は特許出願済み、□は検証済み。なお、テーラーメードモデリング特許出願済