# ビッグデータ時代のデータ解析手法、
# KY (K-step Yard sampling)法の展開

## Data analysis methods in the big data era, Development of the KY (K-step Yard sampling) methods

株式会社　インシリコデータ　湯田　浩太郎

### １．はじめに

現在までに展開されてきた多変量解析／パターン認識は、精度や信頼性の高い解析にはデータ解析手法に応じた適切なサンプル数が必要であり、一般的にはサンプル数が多くなると多変量解析／パターン認識の適用には注意が必要である。

現在はインターネットの普及により日々集積されるデータ量が膨大であり、更にはデータ密度の高い画像データや音声データも簡単、且つ大量に集めることが可能となっている。いわゆる、従来とはデータ量が桁違いに多い「ビッグデータ時代」となっている。

この「ビッグデータ時代」では、極めて大量のデータを扱うことが可能なデータ解析手法が求められる。現時点では、大量のデータを扱うという観点では合格であるが、データ解析精度を犠牲としたトレンド解析的な手法が暫定的に適用されている。今後は、従来の多変量解析／パターン認識のようなデータ解析精度を有しつつ、様々な要因解析も可能となる新たなデータ解析手法の展開が必要である。

湯田は以前、大量のデータを扱うことが出来て、且つ分類精度を極大にする「KY(K-step Yard sampling method)法」を開発し、発表している。この KY 法を用いて Ames 試験データを用いて毒性分類を実施し、当時としては破格の約７０００サンプルを用いて発がん性ポジ／ネガの完全分類を実現した。

KY 法の大量データへの対応力が証明されたので、その後 KY 法の改良を行い、また二クラス分類のみならず、フィッテイング（重回帰）に対応できる KY 法も開発した。本発表では、最近の KY 法の展開と内容についてまとめる。

### ２．討論

#### ２．１KY 法の提案と概要

KY (K-step Yard sampling) 法は既存のデータ解析手法を運用することで、極めて高い分類率を達成するデータ解析手法として 2006 年にニクラス分類手法として提案された。この手法の特徴を以下に列記する。

①二クラス分類 KY 法の特徴（図１）

・サンプルデータを３分割する（クラス１、クラス２およびクラス分け困難）

・クラス分け困難とされたサンプルを対象として、同じ分割操作を繰り返す

・最終的に全サンプルが分割された時点が最終ステップとなる

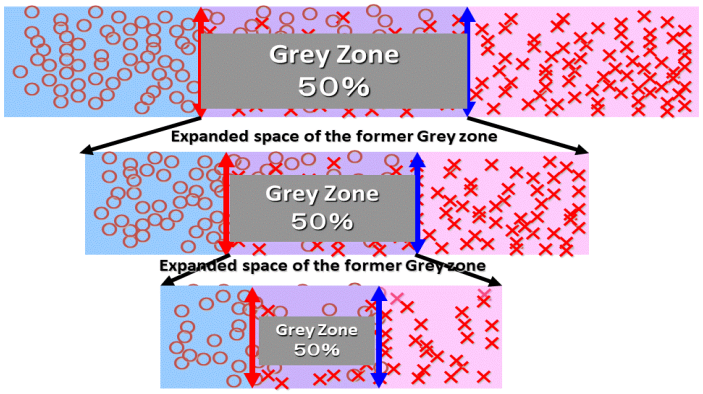


図1．二クラス分類KY(K-step Yard sampling)法

本手法でサンプル群を３分割するための手法は、従来から展開されてきた二クラス分類手法を適用する。このため、KY 法のための特別な二クラス分類手法は存在しない。手法的には従来手法の運用手順を工夫することで、従来手法では達成困難な極めて高い分類率（略 100%）を達成する。従って、KY 法は最近展開されている AdaBoost やランダムフォレストのような、分類手続きを工夫するアンサンブル学習に類似する。KY 法自体は「多段階的分割サンプリング」の基本に基づき、従来手法の分類アルゴリズム運用を工夫したメタ解析手法となる。

②二クラス分類 KY 法適用可能性

・大量のサンプルの扱いが可能（ビッグデータ時代にも対応）

KY 法はサンプル群を一度の計算で分類するのではなく、段階に分けて順に分類する。この結果、高い分類精度を保ちつつ大量のサンプルを精度高く分類することが可能である。

・サンプルを分類可能性に従ってグルーピング可能

KY 法実施途中に分類作業を終了させることで、分類可能なサンプル群と分類困難なサンプル群へとグルーピング出来る。

現在までに二クラス分類 KY 法として新たに二種類の手法が開発され、総数で三種類となる。これらの手法は総て KY 法の特徴となる多段階繰り返し分割手順を取る。

## ２．２KY 法のフィッテイング（重回帰）への拡張展開

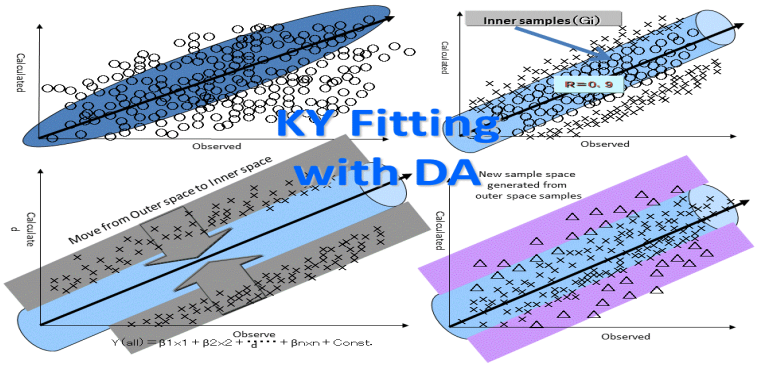


図２．フィッテイング（重回帰）KY法

KY 法は最初に二クラス分類手法として開発された。この KY 法を特徴付ける「多段階的分割サンプリング」の技術を生かして、もう一つの重要なデータ解析手法であるフィッテイング（重回帰）手法への展開を目指した。

①フィッテイング（重回帰）KY 法の特徴（図２）

・サンプルデータを２分割する（残差の小さなグループ、残差の大きなグループ）

・残差の大きなグループを取り出し、同じフィッテイング操作を繰り返す

・最終的に、これ以上回帰式が作成できないサンプル数となった時点で停止する

フィッテイングKY法では、二クラス分類KY法と同様にサンプル群を残差の小さな（良質な）サンプル群と、残差の大きな（ノイズ）サンプル群にグルーピングされる。

現在、フィッテイング（重回帰）KY法として三種類が開発されている。

## ３．まとめ

大量のサンプルを扱いながら解析精度を向上させてゆくことが可能なKY法は、来るべきビッグデータ時代に対応できる手法となる。KY法は当初二クラス分類に対して開発された（２モデルKY法)。その後、同じ二クラス分類でありながら、より簡易なアルゴリズムを採用した二クラス分類KY法（１モデルKY法）が開発された。さらに、予測精度を向上し、且つ自動的に実行する自動プログラムの開発に適したKY法（モデルフリーKY法）を開発した。現在二クラス分類KY法は三種類存在する。

KY法の基本原理を適用しつつデータ解析のフィッテイング（重回帰）に対応するKY法の開発も実施された。このフィッテイング（重回帰）KY法も、アルゴリズムの内容に従って三種類開発された。この結果KY法は二クラス分類及びフィッテイング（重回帰）の双方に対応可能となっている。個々の手法の詳細は発表時に説明する。

**KY法関連発表リスト：**


１．K-step Yard samples法の開発とADME-T予測への適用
○湯田浩太郎　富士通株式会社
第３４回構造－活性相関シンポジウム，Nov.14-15，2006，新潟、K06

２．「完全分類を実現するK－step Yard sampling method (KY法)の提案：
Ames試験データ7000サンプルへの適用実験」
○湯田浩太郎(1,2)　富士通株式会社1、大阪大学　臨床医工学融合研究教育センター2
計算機統計学会第２２回大会 セッション4A-2要旨　2008，Akita

３．A NEW CLASSIFICATION METHOD SUITABLE FOR TOXICITY SCREENING OF CHEMICAL COMPOUNDS
○ Kohtaro Yuta　In Silico Data，Ltd.，
EuroQSAR 2010，19-24　September，Rhodes，Greece

４．毒性予測分野での適用を目指した新規分類/予測手法の提案
○湯田浩太郎　（株式会社　インシリコデータ）
第38回構造－活性相関シンポジウム，Oct. 30-31，2010，徳島、KP20

５．Skin sensitization study by qualitative structure-toxicity relationships (QSTR) by the K-step Yard sampling (KY) method
○Kohtaro Yuta1，Jose M. Ciloy2，Kazuhiro Sato3，Yukinori Kusaka3
第39回構造－活性相関シンポジウム，Nov. 28-29，2011，野田、KP12

６．「KY(K-step Yard sampling）法の展開（二クラス分類および重回帰)」
湯田　浩太郎　株式会社　インシリコデータ
日本計算機統計学会 第 26 回シンポジウム プログラム　セッション2A-1 2012，Tokyo

７．NEW APPROACH FOR QSAR AND QSTR TREND ANALYSIS ON LARGE SAMPLE DATA SET BY THE KY-METHODS

○ Kohtaro Yuta In Silico Data, Ltd.,

19th EuroQSAR 2012, 26-31 August, Vienna, Austria

８．Development of the KY-methods for use on toxicity prediction

○ Kohtaro Yuta In Silico Data, Ltd.,

Eurotox 2013, 1-4 September 2013, Interlaken, Switzerland

９．安全性予測をターゲットとした２クラス分類 KY 法の改良と新規開発

○湯田　浩太郎　株式会社　インシリコデータ

第 41 回構造ー活性相関シンポジウム, Nov. 7-8, 2013, 西宮、KP13

１０．SKIN SENSITIZATION STUDY FROM ONLY ANIMAL DATA BY QUALITATIVE STRUCTURE-TOXICITY RELATIONSHIPS (QSTR) APPROACH

Kazuhiro Sato1, Kohtaro Yuta2, Yukinori Kusaka1

1Department of Environmental Health, School of Medicine, University of Fukui, Fukui910-1193, Japan 2In Silico Data o Ltd, Narashino, Chiba275-0025, Japan

Eurotox 2014, 2014,

１１．Development of in silico (computational) toxicity screening methods

K.Yuta, In Silico Data, Ltd. (Japan) WC9, 2014, 8-25-28, posterII-3-436

１２．「ビッグデータ対応の二クラス分類:KY (K-step Yard sampling methods) 法の開発と展開」

湯田　浩太郎　株式会社　インシリコデータ

2014 年度　統計関連学会連合大会, 2014, 6-10, Tokyo

１３．New approach for QSAR and QSTR trend analysis on large sample data set by the KY-methods

○ Kohtaro Yuta In Silico Data, Ltd., Eurotox 2015, Poster P06-010,

１４．Development of new data analysis methods: KY-methods

K.Yuta, In Silico Data, Ltd. (Japan)

WC10, 2017, 8-20-24, Seattle, U.S.A., posterII-3-436

１５．Development of a state-of-the-art multiple regression KY-method corresponding to the big data era and its application to fish toxicity

○ Kohtaro Yuta In Silico Data, Ltd., Eurotox 2018, Poster P05-26, Brussels, Belgium

１６． Chemical Data Science for Drug Design:

New Development of the KY-methods in Big Data Era

K.Yuta, In Silico Data, Ltd. (Japan)

5th International Symposium of Medicinal Sciences, Poster;IP-12, 2019, Chiba, Japan