常に完全（100%）分類を実現する KY 法の開発（K-step Yard sampling methods）

Development of the KY Methods which Achieves Perfect (100%) Classification under Any Conditions

湯田 浩太郎[a]

Kohtaro Yuta

1. はじめに

現在，効率的な化合物開発技術としてインシリコ[1]による様々なスクリーニング手法が開発され，時代の発展とともにスクリーニング対象は薬理活性および物性から安全性（毒性）へとシフトしつつあります。安全性スクリーニングでは，創薬研究で行われるドッキングによる薬理活性スクリーニングの適用は基本原理上出来ません。代わりに，多変量解析／パターン認識によるインシリコスクリーニングが実施されます。

2. インシリコでの安全性（毒性）評価の困難性

インシリコによる化合物安全性評価の特徴は，①評価対象となる化合物の種類が多種多様である（結果として，化合物構造変化性が極めて高い），②取り扱うサンプル数が多い，③極めて高い予測率と信頼性が要求される，という 3 項目につきます。線形および非線形手法の差を問わず，従来の多変量解析／パターン認識手法を用いてこれらの 3 項目が内包する問題点を解決することは殆ど不可能です。

3. KY 法の開発と，完全分類の実現

KY 法（K-step Yard sampling methods）[2]は，インシリコ上での安全性評価で大きな障壁となる，前記 3 項目のクリアを目的として開発された新しいデータ解析手法です。なお，KY 法は現時点で二クラス分類手法として 3 種類，重回帰のようなフィッティング手法として 3 種類の手法が展開されています[3]。

KY 法自体は新しいのですが，内部で利用される分類手法（データ解析手法）は従来の手法を用いて行いますので「メタ解析手法」となります。つまり，KY 法で利用される判別関数は従来手法を用いて構築されます。KY 法では，サンプル数が極めて大きい場合や，サンプル群の重なりや分散が極めて高い場合であっても常に完全分類，あるいは極めて高い相関係数（R）及び絶対係数（R2）が実現されます。

4. 二本の判別関数を用いた二クラス分類

二クラス分類に用いられる KY 法として 3 種類ありますが，本稿では 2 本の判別関数を用いて分類／予測を行う KY 法（2 モデル KY）について説明します。

この 2 モデル KY を行う大まかな手順を以下に示します。判別関数作成等の細かな操作は US 特許[2]をご参照ください。

手順 1；サンプル空間をポジサンプル（Figure 1 中◯）のみ，およびネガサンプル（Figure 1 中×）のみ存在するゾーン（空間）と，ポジとネガが混在するグレーゾーンに分割する。

手順 2；グレーゾーンに帰属するサンプル群を初期サンプルセットとし，再び手順 1 の操作により 3 ゾーンに分割する。

手順 3；上記の手順 1 と手順 2 の操作を繰り返し，最終的にグレーゾーンのサンプルが無くなった時点（1 本の判別関数で分類完了）で計算を終了する。この時点で完全分類が実現されたことになります。

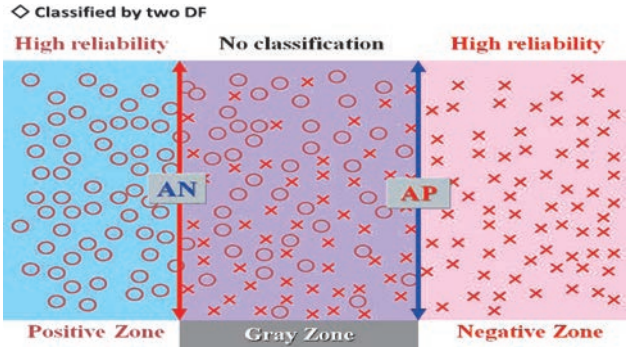

Figure 1. Making three zones by combination of two different discriminant functions.

Figure 1 中，AN（All Negative）の赤い線は，ネガサンプル（図中×）を 100%（完全）分類する判別関数です。AP（All Positive）の青い線は，ポジサンプル（図中◯）を 100%（完全）分類する判別関数です。AN と AP の二本の

[a] 株式会社　インシリコデータ
連絡先　〒 275-0025　千葉県習志野市秋津 5-19-5
電子メール　contact@insilicodata.com

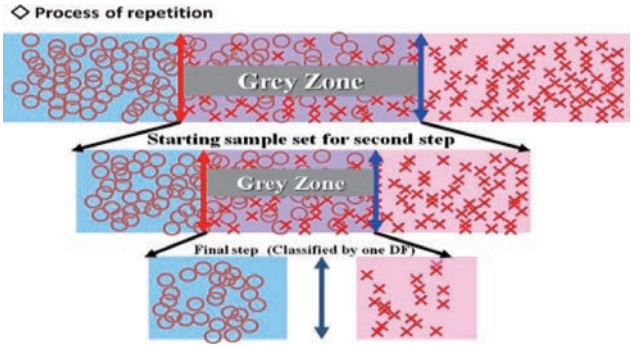

Figure 2. Repetition calculation using the grey zone sample by the KY methods.

判別関数に挟まれた真ん中の領域はポジ／ネガサンプルが混在するグレーゾーンとなります。これが手順 1 の操作であり，これで KY 法による二クラス分類の第 1 ステップが完了し，手順 2 に進みます。

手順 2 では，第 1 ステップでグレーゾーンに帰属されたサンプル群を第 2 ステップの初期サンプルセットとします。手順 1 と同様の手順（特徴抽出等の実施）を実行し，新たなサンプル空間を作り直します。再び AN および AP の判別関数を作成し直し，第 1 ステップ同様にポジ／ネガおよびグレーゾーンに分類し，第 2 ステップを完了します。

この一連の手順を繰り返し，最終的にサンプルが一本の判別関数（Figure 2 では最下層の段階（第 3 ステップ））で完全分類出来た時点を最終ステップとし，分類を完了します（手順 3）。これで，用いた全サンプルの完全分類が実現します。(Figure 2)

5. KY 法の特徴と機能／効果

二クラス分類での KY 法の二大特徴を以下に示します。

特徴 1. サンプル空間をサンプルの分布状態に従って 3 領域（ポジ領域，ネガ領域，ポジ／ネガ混在領域（グレーゾーン））に分割する。

特徴 2. ポジ／ネガ混在領域（Figure 1，2 中ではグレーゾーン）を再分類し，この混在領域が無くなるまで繰り返す。

上記二つの特徴を有する手法が KY 法となります。現在開発されている二クラス分類用の 3 種類の KY 法は総て上記の二特徴を満たします。この特徴 1 と特徴 2 の操作を行うことで，KY 法は他の手法には無い以下の優れた機能を有するデータ解析手法となります。

①**サンプル数に関係なく**常に完全分類実現

②**クラス間重複の高いサンプル群**も完全分類実現

上記のように，サンプル数やサンプル間の重なり度に関係なく常に完全（100%）分類を実現する強力な分類能力は，第二節で述べた，安全性（毒性）をスクリーニングする時に要求される三項目の要求事項を総て満たします。

6. KY 法による Ames 試験データ（6965 化合物）の完全分類実現

KY 法の強力な分類能力を証明する実験として，化合物の変異原性評価試験法である Ames 試験のデータ（6965 化合物（Mutagen:2932, Non-mutagen:4033））を用いて解析を実施しました[4]。解析対象となる化合物群はメタン／エタンレベルの小さな化合物群からテルペン，ステロイド，糖類，マクロライド等の多種多様の化合物より構成されます。結果として化合物の構造変化性が非常に高く，この点で二クラス分類は極めて実施困難です。さらに，サンプル数も約 7000 という大きな数であり，従来手法による完全（100%）分類の実現は殆ど不可能で，極めて扱いの難しいサンプルセットとなります。

KY 法の適用により，前記 6965 の全サンプルは 23 ステップで完全に分類出来ました。最後の 23 ステップ目は一本の判別関数のみを用いて分類しております。なお，本解析に用いたソフトウエアは ADMEWORKS の ModelBuilder[5] です。

Table 1. Perfect classification of the Ames test sample (6965) by the KY methods.

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

7. KY 法のまとめと今後

KY 法は従来手法では実現できなかった強力な分類能力を有します。サンプル数がどんなに多くとも，またクラス間重なりが極めて大きいサンプル群であっても常に完全（100%）分類が実現可能となります。

KY 法中で利用される分類アルゴリズム（手法）は総て従来手法を用いており，従って KY 法は「メタ解析手法」となります。例えば，KY 法の二クラス分類で用いられる AP および AN の判別関数は，従来から展開されている Bayes 判別分析，ニューラルネットワーク（NN），サポートベクターマシン（SVM），さらには AdaBoost 等を利用して作成可能です。

KY 法は「メタ解析手法」なので，基本原理を理解されれば従来手法の多変量解析／パターン認識ソフトを用いても実施可能です。実際の細かな手順は US 特許[2] に記載されて

いますので，これを参考にして KY 法を実施いただければと存じます。

先の ADMEWORKS の ModelBuilder[5] を用いれば，KY 法の手順や予測モデルの構築，さらには PREDICTOR を用いた KY 法の予測モデルを用いたイントラネットワーク上での化合物スクリーニング等を総合的，かつ簡単に実施できます。

引用文献

（1） http://ja.wikipedia.org/wiki/In_silico

（2） Yuta, K. U.S. Patent 7 725 413, 2010.

（3） http://insilicodata.com/themas/Patent%20table.html

（4） 湯田浩太郎，第 34 回構造活性相関シンポジウム，富山，2006 年 11 月 14–15 日，K06.

（5） http://jp.fujitsu.com/group/kyushu/services/lifescience/admeworks/index.html

（受理日　2012 年 2 月 24 日）