ＣＢＩ学会和文誌、2 巻、1-5 ページ、2009 年　　解説論文

機械学習による薬物のヒト肝ミクロゾーム安定性の予測

嵜山陽二郎*
Pharmacokinetics Dynamics Metabolism, Pfizer Global R&D, Sandwich Laboratories

幸瞳、本間光貴
理化学研究所生命分子システム基盤研究領域・制御分子設計研究チーム

***Sandwich, Kent CT13 9NJ, UK**
E-mail: Yojiro.sakiyama@pfizer.com

(論文受付日 February 6,2009 ; 公開日 February 26,2009)

要旨：　薬物の吸収、分布、代謝および排泄（ADME: absorption, distribution, metabolism and excretion）などの薬物動態の指標の測定値（エンドポイント）は、多くの場合説明変数に対して非線形であり、それらを予測するために機械学習（machine learning）の手法など工学領域に端を発する高度な解析技術が必要とされるようになった。我々は、肝ミクロゾーム安定性試験データを材料に、最近機械学習手法として広く利用されている６種類の分類器を用い安定性の予測を行った。その結果、ランダムフォレスト、サポートベクターマシン、ガウス過程法などの最近の非線形手法が、旧来の分類器に比べ優れた性能を発揮することが示唆された。今後もこれらの手法の幅広い応用が期待される。

キーワード: ADME、QSMR, 薬物動態、機械学習、ミクロゾーム、ランダムフォレスト、サポートベクターマシン、ガウス過程法

１．はじめに

薬物動態（ADME）をコンピューターを使って予測することは、予算のかかる ADME 試験の節約と効率化の観点で重要である。薬物の代謝エンドポイントを化合物情報を使って予測する代表的方法としては QSMR （ quantitative structure-metabolism relationship; 定量的構造代謝相関）がある。その基本的枠組みは、QSAR（定量的構造活性相関）と同様で、化合物情報を記述子（descriptors）として数量化し、これを入力変数として化合物の代謝エンドポイントを予測するというものである。QSMR アプローチでは一般に訓練データとする化合物の数および入力変数（記述子）の数が多く必要とされる。また代謝エンドポイントの挙動も多くの場合非線形である。したがって、線形かつ少ない変数を主体とした旧来のデータ解析ソフトウェアを用いて予測する上では限界がある。一方、人工知能などの工学領域において機械学習（machine learning）の手法が最近十数年の間急速に発展し、非線形問題に有効なカーネル法や、ロバストな集団学習アルゴリズムなど、多くの優れた予測手法がケモインフォマテ

ィックスの領域にも徐々に浸透してきている。またその解析パッケージもウェブ上でフリーソフトとして公開され、誰でも手持ちの PC で解析出来るようになった。

そこで今回我々は、フリーソフトウェア R および WEKA を用いた肝ミクロゾーム分画における化合物の in vitro での安定性の予測手法について紹介することとする。

２．材料

２．１　肝ミクロゾーム安定性試験

被験化合物(1μM)を、ヒト肝ミクロゾーム(0.78mg protein/ml)、NADPH、MgCl2 の存在下、リン酸カルシウム緩衝液中 37℃でインキュベーションし、経時的に採取した試料中の被験化合物を HPLC-MS で定量した。各時点のデータより回帰的に算出した消失半減期をもとに、in vitro 固有クリアランス（CLint, ml/min/kg）を算出した[1]。この値は化合物の第１相の代謝に対する安定性の指標として広く使われている。

２．２　データおよび入力変数

ファイザー化合物ライブラリーより肝ミクロゾーム安定性試験済みの 2,439 個化合物のデータを用意した。これらを Clint の区間別に(テスト化合物、訓練化合物それぞれについて各区間の Clint の割合が等しく偏りがないように) 無作為抽出を行い、487 個をテスト化合物、1,952 個を訓練化合物とした。カットオフ値（ｆ）を設定し、これらの化合物のうち、CLint（ml/min/kg）＜ｆ の化合物は“代謝安定”化合物、CLint ≧ｆの化合物を“代謝不安定”化合物とした。入力変数には、MOE2005.06（Molecular Operating Environment）により計算した 193 個の 2 次元記述子を用いた[1]。

３．解析方法

３．１　ソフトウェア

R は２千を超えるフリーパッケージを所有するオープンソースのデータ解析・データマイニング専用ソフトウェアである。R プロジェクトの中核をなす CRAN のホームページには各機械学習手法の理論的詳細の他、フィッシャーのあやめなど簡易データによる解析事例が収載されている。また、既に多くの R に関する書籍も出版されており、これらを参考にすれば初心者でも手軽にプログラムを組み解析が可能である[2]。

オープンソース型のソフトウェアとしては他に WEKA が広く用いられている。WEKA は機械学習の専門家よりはむしろエンドユーザーに便利な設計となっている[3]。

３．２　いろいろな機械学習

機械学習とは端的に言えばコンピューターを使ってルールを抽出する手法である。現在、百種類以上の機械学習の手法が開発されているが、これらは大きく分けると教師あり学習（supervised learning）、教師なし学習（unsupervised learning）および強化学習（reinforcement learning）の３種類に大別される。教師なし学習は主成分分析に代表されるような、入力変数のみからルールを抽出する手法である。強化学習は迷路学習のような試行錯誤によりルールを発見する手法である。最も広く用いられているのは教師あり学習で、これは入出力のペアからなるデータからルールを抽出し、そのルールをもとにある入力から未知の出力を予測する方法である。さらに予測する目的変数が連続変数の場合は回帰（regression）、カテゴリー変数の場合の分類（classification）と分けられる。今回我々は後者の分類のツールとして、最近広く使われている 6 種類の分類器をとりあげ、これらについて以下に概説する。

３．３　ナイーブベイズ分類器

入力変数 $\mathbf{x}=[x_1,\ldots x_m]$ が与えられたとき、あるクラス c_i に属する確率は、ベイズの定理により、$P(c_i|\mathbf{x})=P(c_i)\cdot P(\mathbf{x}|c_i)/P(\mathbf{x})$ となる。ナイーブベイズ分類器は、$x_1,\ldots x_m$ が独立であるという仮定のもとで、$P(c_i)/P(\mathbf{x})=k_i$（定数）とすると、各クラスに属する確率は単純（ナイーブ）に与えられたクラスにおける各 $x_1,\ldots x_m$ の生起確率の積で表され、$P(c_i|\mathbf{x})=k_i\cdot\prod_j P(x_j|c_i)$ となる。最終的に $P(c_i|\mathbf{x})$ の最も大きいクラスに属すると判定する。R ではパッケージ klaR または e1071 を用いる。

３．４　決定木による分類

決定木による分類は、親データのテストを枝の分岐点（node）で行いその結果に応じて次の枝先の分岐へと子データが振り分けられる。その子データが今度は親となり再帰的に次の分岐へと進む。最終的

に木の末端の葉に、結果のクラスが割り振られる。分岐の基準としては、分岐前と分岐後とでデータ内のクラスのばらつきがどれだけ減少したかを定量化する基準が用いられる。CART（classification and regression tree）では分岐の基準としてジニ係数もしくはエントロピーを用いる（ここではジニ係数を用いた）。R ではパッケージ rpart または mvpart を用いる。

３．５　ランダムフォレスト

ランダムフォレストは、決定木による集団学習（ensemble learning）手法、すなわち複数の決定木による学習結果を統合し、精度を向上させる学習手法である。先ず親データから復元抽出（bootstrapping）により複数の子データが作成される。このとき変数もランダムに抽出される。それぞれの子データから得られた決定木分類の結果を統合（回帰では平均、分類では多数決）し最終結果を得る。理論的詳細は[4]を参照。Ｒではパッケージ randomForest を用いる。

３．６　サポートベクターマシン

サポートベクターマシン（SVM)は、高次元の特徴空間で線形分離を行なう解析手法である。学習の結果得られた識別境界（超平面）においては、境界に最も近いサンプルとの距離（マージン）が最大化される。識別が完全でない場合も、スラック変数の導入により最適化される。

SVM はカーネル法の一種で、線形分離不可能な問題に対しても、カーネル関数を用いて特徴空間へ写像し、特徴空間上で線形分離を行うことができる。なお実際の解析では内積計算の部分をカーネル関数に置き換えるにすぎない（カーネルトリック）。

SVM のような２次計画問題は局所解の問題が回避できるという長所があるが、そのままでは計算時間がかかるという短所もある。計算効率を高める手法として SMO（sequential minimal optimization）が広く利用されている。理論的詳細は[5]を参照。R ではパッケージ e1071 または kernlab を用いる。

３．７　ガウス過程分類法

ガウス過程法は、正規分布にしたがう関数群の事前分布を仮定し、ベイズ推定の下で事後確率を計算し予測値を得る解析手法で、非線形問題に対し頑健な手法として最近注目されている。いま、n 個の入力変数行列 $\mathbf{X}=[\mathbf{x}_1,\cdots\mathbf{x}_N]$ と n 個の観測値 $\mathbf{y}=[y_1,\cdots y_n]^T$ が与えられたとする。ガウス過程では、入力変数ベクトル $\mathbf{x}_i$ の関数 $f(\mathbf{x}_i)$ を想定し、その関数群 $\mathbf{f}=[f_1\cdots f_n]^T$ が入力 $\mathbf{X}$ と超パラメータ θ のもとで正規分布をする仮定する。すなわち、$\mathbf{f}|\mathbf{X},\boldsymbol{\theta}\sim\mathbf{N}(\mathbf{0},\mathbf{K})$。ここで $\mathbf{K}$ は共分散関数 $k(x_i,x_j)$ で構成された $\mathbf{n}\times\mathbf{n}$ の共分散行列である。

２値分類問題では、解析上の便宜のため、関数 $f(\mathbf{x}_i)$ はシグモイド関数 $\Phi(a)$ を用いて単位区間上に写像し、$p(y=1|\mathbf{x}_i)=\Phi(f(\mathbf{x}_i))$ とする。ここで、$p(\mathbf{y}|\mathbf{f})$ はベルヌーイ分布にしたがい、$p(\mathbf{y}|\mathbf{f})=\prod_{i=1}^{N}p(y_i|f_i)=\prod_{i=1}^{N}\Phi(y_i f_i)$ となる。ベイズの定理により、$\mathbf{f}$ に対する事後分布は以下となる。

$$p(\mathbf{f}|\mathbf{D},\boldsymbol{\theta})=\frac{N(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{p(\mathbf{D}|\boldsymbol{\theta})}p(\mathbf{y}|\mathbf{f})=\frac{N(\mathbf{f}|\mathbf{0},\mathbf{K})}{p(\mathbf{D}|\boldsymbol{\theta})}\prod_{i=1}^{N}\Phi(y_i f_i)$$

周辺化により、入力 $\mathbf{x}^*$ に対し予測したいクラスラベル y^* は以下で表される。

$$p(y^*|\mathbf{D},\boldsymbol{\theta},\mathbf{x}^*)=\int p(y^*|f^*)p(f^*|\mathbf{D},\boldsymbol{\theta},\mathbf{x}^*)df^*$$

さらに第２項は以下で表される。

$$p(f^*|\mathbf{D},\boldsymbol{\theta},\mathbf{x}^*)=\int p(f^*|\mathbf{f},\mathbf{X},\boldsymbol{\theta},\mathbf{x}^*)p(\mathbf{f}|\mathbf{D},\boldsymbol{\theta})d\mathbf{f}$$

ここで、$p(\mathbf{f}|\mathbf{D},\boldsymbol{\theta})$ は解析的に解けず、最終解を得るための様々な近似手法が提案されている。理論的詳細は[6]を参照。R ではパッケージ kernlab を用いる。

３．８　k 最近傍法

k 最近傍法は判別すべき個体の周辺の最も近いものを k 個見つけ、その k 個の多数決により、どのグループに属するかを決める方法である。距離の測度としては一般にユークリッド距離が用いられる。k 最近傍法は学習ツールというよりむしろ、データの記憶に基づいて判別を行う記憶ベース推論法である。R ではパッケージ class を用いる。

４．性能評価

４．１　クロスバリデーション（交差検証法）

一般に多くの入力変数を使った複雑な非線形モデルでは、訓練データそのものに対してはほぼ完璧な予測が得られるが、ここには過学習（オーバーフィッティング）の問題がある。モデルは本来訓練データに含まれない新規データに対して高い予測性能（汎化性能）を示すことが望ましい。そこで、データが十分にあるときは、その一部を抽出し独立し

た検証用データ（validation set）を作成し、これをモデルの評価に用いる。データが十分にない場合、もしくは出来るだけ多くのデータをモデル作成に使いたい場合は、交差検証法が有効である。交差検証法では、訓練データを N 個に分割し、そのうち N-1 分画でモデルを作成しこれを用いて残りの 1 分画を予測する。これを N 個の分画すべてについて N 回実施し、これらを統合して最終結果とする。このようにすると予測対象のデータは常にモデルに含まれないため、モデルの汎化性能を評価することができる。これは LGO（leave group out）法と呼ばれ、特に N がデータ数と等しい場合は LOO（leave one out）法と呼ばれる。LOO 法はデータ数が少ない場合に有効な方法であるがデータが偏りやすくまた計算時間がかかるなどの問題がある。経験的には、N=10 が良いことが知られている[7]。筆者らは N=10 として（10-fold cross validation）実施した。

４．２　評価の指標

分類器の性能評価は、分類器による予測結果が実際の観測結果とどの程度一致するかに基づいて行われる。いま、目的変数がある閾値を越える場合（陽性）もしくは越えない場合（陰性）を予測したいとしよう。陽性観測値を陽性と予測する確率（真陽性確率）は感度（sensitivity）、陰性観測値を陰性と予測する確率（真陰性確率）は特異度（specificity）と呼ばれる。全体として観測と予測が一致する確率は、一致度（accuracy）と呼ばれる。単に正しい予測の割合を知りたい場合には一致度は便利な指標である。また、これをさらに改良したものとして、偶然による一致を調整した Kappa 係数、各クラスのデータサイズのアンバランスを調整した MCC（Matthews correlation coefficient）[8]、各クラスのデータサイズそのものを調整した Youden 係数[9]などがある。これらはすべて、ある特定の閾値下での予測性能の指標である。

一方、これらの指標は閾値の変化に伴って変動する。極端な例では、閾値を非常に高い値に設定すると、すべての観測値は陰性と予測され、陽性と誤って判定する偽陽性の確率は０となるが、この場合予測そのものが無意味となり決して良い分類器とはいえない。したがって予測性能は、常に閾値との関係に基づいて評価することが望ましい。そのための便利なツールとして、ROC 曲線（receiver operating characteristics curve）がある。ROC 曲線は、真陽性確率（感度）と偽陽性確率（1-特異度）の関係をプロットしたものである。様々な閾値でこれらの関係をプロットすると、図１に示すように、左上に凸の曲線が描かれる。点の位置は閾値が小さいほど原点に近く、大きいほど原点から離れる。どの閾値を用いるかは研究目的により異なるが、スクリーニングの初期で予測ミスによる有望化合物の棄却を回避したいのであれば閾値を低めに設定する保守的な措置がとられる。

ROC 曲線はまた分類器の性能評価においても優れたツールである。ROC 曲線における最も左上の点（感度と特異度がともに１）は観測と予測が完全に一致する最も理想的な点で、曲線がこの点に近いほど良い分類器となる。言い換えれば、曲線下面積（AUC: Area Under the Curve）が大きいほど良い分類器であり、曲線下面積が、閾値に関係ない全体としての分類器の性能を反映する。

５．解析結果

図１に、肝ミクロゾーム安定性試験データについて、6 種類の分類器を用いて得られた ROC 曲線を示す。曲線上の各点は、種々の閾値（CLint, 7,10,20,50,100 ml/min/kg）において各分類器を用いて 10-fold cross validation により予測した結果から、真陽性確率（感度）と偽陽性確率（1-特異度）を算出しこれらをプロットしたものである。それぞれの曲線下面積（ROC_AUC）を算出した結果、表１のように、大きい順に RF > SVM > GP > kNN > CART > NBC　となった。

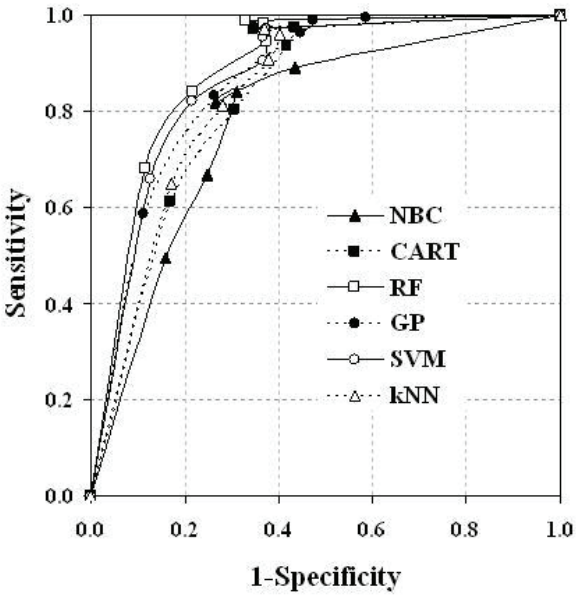

図 1　ROC 曲線：肝ミクロゾーム安定性試験データについて 6 種類の分類器を用い予測した。NBC: naïve Bayes classifier　CART: classification and regression tree　RF: random forest　GP: Gaussian Process　SVM: support vector

machine　kNN: k nearest neighbor

	Accuracy	Kappa	MCC	Youden	AUC_ROC
NBC	0.801	0.477	0.480	0.455	0.760
CART	0.839	0.562	0.572	0.519	0.820
RF	0.858	0.616	0.625	0.572	0.880
GP	0.849	0.577	0.599	0.515	0.856
SVM	0.844	0.594	0.596	0.574	0.862
kNN	0.839	0.576	0.580	0.548	0.837

表 1　6 種類の分類器で得られた結果をもとに算出した各種予測性能の指標

また表１では、6 種類の分類器を用いて得られた予測性能について、各種予測指標による比較を行った。先ず一致度では、どの分類器も 0.85 前後の値であり、この結果に関しては 8 割強は正しく予測していると考えられる。

ここで、今回用いた 6 種類の分類器のうちどれが最も優れた分類器であるかを判定したいとする。すると、Youden 係数は SVM の方が RF よりやや高い値であるが、それ以外の指標については RF の方が高く、これらの値だけでは SVM と RF のいずれが良いか判断し難い。しかし、先に述べたように、一致度、Kappa、MCC および Youden 係数はすべてある特定の閾値下での予測性能の指標であるのに対し、ROC 曲線下面積は様々な閾値における予測結果を総合的に反映した指標である。したがって、このような場合は ROC 曲線下面積による判断が妥当であり、RF が最も良い分類器であると判断することができる。

６．考察

今回得られた結果から、集団学習アルゴリズムやカーネルマシンなど最近の非線形分類器は、決定木分析などの旧来の手法に比べ予測性能が高く、なかでも Random Forest は最も高い性能を有することが示唆された。また、最近注目されているガウス過程法についても、これらの分類器に匹敵する優れた予測性能を有することが示唆された。筆者らが先に行った４種類のデータによる網羅的解析結果でもこれらと同様の傾向が認められ、それはデータセットの大きさに依存しないことが示唆されている[10]。また、予測性能を正しく評価する上で、交差検証法や ROC 曲線などの優れた評価ツールの利用も望まれる。また並行して、特徴選択（feature selection）によるモデルの簡素化、domain of applicability や chance correlation などの評価も併せて進めることにより、一層優れたモデル構築・モデル選択が可能となるであろう。

最近の非線形分類器の性能の良さはバイオインフォマティクスなど他の研究領域でも既に知られているが、一方ではパラメータチューニングや計算時間の負担などの問題もあり、これらへの取り組みは工学領域において進められている。今後も機械学習手法の薬物動態予測への幅広い応用が期待されるが、工学領域など他領域の研究者とのインタラクションにより本研究領域がより一層活性化することが期待される。

参考文献

[1] Sakiyama Y, Yuki H, Moriya T, et al. Predicting human liver microsomal stability with machine learning techniques. *J Mol Graph Model*, **26**,907-15(2008).

[2] Team RDC. R. *A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Australia*. http://www.R-project.org. (2005).

[3] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. http://www.cs.waikato.ac.nz/~ml/weka/.

[4] Breiman L. Random forests. *Machine Learning*, **45**,5-32(2001).

[5] Cristianini N, Shawe-Taylor J. *An introduction to Support Vector Machines*. Cambridge University Press, New York, (2000).

[6] Rasmussen CE, Williams CKI. *Gaussian processes for machine learning*. The MIT Press, (2006).

[7] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (Morgan Kaufmann, San Mateo),* **2**,1137-43(1995).

[8] Baldi P et al. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**,412-24(2000).

[9] Youden WJ. Index for rating diagnostic tests. *Cancer*, **3**,32-35(1950).

[10] Sakiyama Y. The use of machine learning and nonlinear statistical tools for ADME prediction. *Expert Opin Drug Metab Toxicol* 2009, in press.