Vol.2009-BIO-17 No.10
2009/5/25

ブースティングによる薬物クリアランス経路予測

池田和史[†1] 年本広太[†1] 草間真紀子[†2]
前田和哉[†2] 杉山雄一[†2] 秋山泰[†1]

薬物のクリアランス経路を特定することは薬物動態学における重要な課題である．そこで，本研究では教師あり機械学習の手法を用いて，既知の薬物の物理化学的特性から薬物の主要なクリアランス経路を予測した．先行研究では，解釈性の高い矩形領域法と判別性能の高いサポートベクターマシン (SVM) が用いられていたが，本研究では Boosting のアルゴリズムを用いることで解釈性と性能が両立できるような予測システムの構築を目指した．実装した予測システムを用いて予測実験を行い，その結果を先行研究と比較し，評価した．結果として，本予測システムを用い，SVM に匹敵する汎化性能を持ち，解釈性にも比較的優れた学習を行うことができた．

Prediction of Drug Clearance Pathway by Boosting Algorithm

Kazushi Ikeda,[†1] Kouta Toshimoto,[†1]
Makiko Kusama,[†2] Kazuya Maeda,[†2]
Yuichi Sugiyama[†2] and Yutaka Akiyama[†1]

It is an important problem in pharmacokinetics to determine clearance pathway of drugs. We have developed a prediction system for major clearance pathway of drugs from physicochemical characteristics of drugs. In previous studies, we have proposed a SVM-based system which has superior prediction performance, and another system based on original Rectangular method which shows good interpretationability. In this study, we aim to build a prediction system which shows both good performance and good interpretation, by using boosting algorithm. As a result, new system showed a superior performance almost comparable to SVM, and better interpretation.

1. はじめに

近年，医薬品開発には数百億円の研究費と十年以上の期間が必要とされている[1)]．これには，創薬初期の in vitro 実験系を用いた候補化合物のスクリーニングの難しさや臨床試験における副作用の判明などが理由として挙げられる．また，年々増加する開発費に対し，認可されて市場に出る新薬は毎年 20 個前後で頭打ちの状況であり，新薬開発の生産性が大きく低下している事実がある．現在，このような創薬の現場で薬物の血中濃度推移，臓器への分布特性をはじめとする薬物動態特性の解析が注目を集めている[2)]．

薬物の体内動態を予測する上で重要な情報の一つとして，クリアランス経路が挙げられる．クリアランス経路とは，体内に取り込まれた薬物がどの組織で代謝，排泄されるかを表したものであり，体内の薬物解毒システムにおける重要な情報のひとつである．様々な代謝酵素やトランスポーター群が薬物のクリアランスに大きく関わっているが，薬物の解毒は薬物代謝酵素による化学構造の変更による解毒と，トランスポーターによる細胞の取り込み，そして排除による解毒が巧みにリンクした生体防御機構ととらえることができる．近年，これらの各要素を個々に実測し，その後統合するという方法論による予測法が開発されてきている[2)]．

そこで本研究では，薬物の開発初期段階で得られる基本的な物理化学的特性からクリアランス経路の予測を機械学習の手法を用いて試みた．まず，主要な 5 つのクリアランス経路に対して，クリアランス経路ごとに機械学習の手法を用いた予測を行い，最終的に薬物がどのクリアランス経路にふさわしいかを予測した．また，本研究には先行研究として，サポートベクターマシンによる予測[3)4)] と矩形領域法による予測[4)5)] があり，前者は高速，かつ高精度でクリアランス経路を予測できるが，判別境界が非線形であり，判別の意味が主な利用者である薬学の専門家には不明瞭であった．また，後者は解釈性の高い判別領域を生成できるが，精度があまり良くない，および計算時間が大きいという欠点があった．そこで本研究では，矩形領域法で得られるような解釈性の高い判別領域を生成した上で，高速，高精度を維持できるような予測方法を Boosting[6)] のアルゴリズムを用いて実装することを目指した．

†1 東京工業大学 大学院情報理工学研究科
Graduate School of Information Science and Engineer, Tokyo Institute of Technology
†2 東京大学 大学院薬学系研究科
Graduate School of Pharmaceutical Sciences, The University of Tokyo

Vol.2009-BIO-17 No.10
2009/5/25

2. クリアランス経路

本研究では，主要な経路といえる 5 種類のクリアランス経路，すなわち 3 種類の **cytochrome P450（CYP3A4， CYP2C9， CYP2D6)** を介した代謝，**OATP（Organic Anion Transporting Polypeptide)** を介した肝取り込み，**腎排泄（Renal）**を予測対象とした．

cytochrome P450（以下，CYP) は微生物から植物，動物まで生物界に広く分布する一群のヘムタンパク質である．CYP には，触媒する反応の基質特異性が異なる多数の分子種の存在が知られているが，これらは生命進化の過程で分化した遺伝子ファミリーであると結論されている[7]．本実験では，主に肝臓での代謝を行う CYP3A4， CYP2C9， CYP2D6 の 3 種類をクリアランス経路の候補に用いた．これら 3 つの酵素は薬物代謝のおよそ 8 割を担っている CYP の中でも特にヒトの肝臓内の代謝を司る代表的な酵素であり，CYP3A4 は抗生物質や免疫抑制剤，CYP2C9 は血糖降下薬，CYP2D6 は抗精神薬や抗うつ薬などの重要な薬物の代謝を行っている．

トランスポーターはチャネルやレセプターと共に細胞膜に存在する膜タンパク質の一種であり，細胞の内外の物質輸送をコントロールする働きをしている．薬物が体内を移動する際，トランスポーターの働きがなければ，目標の部位に到達できない．トランスポーターはその機能と性質から様々なファミリーを持っているが，本研究ではその中でも薬物を肝臓に移行させる重要な機能をもった有機アニオントランスポーター OATP（Organic anion transporting polypeptide) ファミリーを予測するクリアランス経路の対象とした．

薬物がヒトの生体内で代謝，排泄される過程で，腎臓による排泄は他の排泄に比べ，無視できない過程である．腎排泄は，糸球体濾過，尿細管における能動的輸送，受動的な尿細管再吸収の 3 過程からなる．これらの過程で非結合型薬物をはじめとした様々な薬物が腎臓によって排泄される．なお，本研究における腎排泄とは，薬物が未変化の状態で腎臓から排泄される場合を指す．

3. 薬物データの処理

本研究の実験に使用した化合物のデータセットは，市場に流通している医薬品に関するものであり，データの総数は 141 化合物である[5]．各データはそれぞれ，**電荷（Charge)，血漿中タンパク質非結合率（fu)，分子量（Molecular Wight;MW)，n-オクタノール/水 分配係数（logD)** の 4 つの特徴量と正解となるクリアランス経路の情報を加えた計 5 つのパラメータを持っている．実験を行う前に，これらのデータを電荷に関して分類する．正の電荷をもつ化合物と中性の化合物に対してはデータ集合 S^+ に，負の電荷をもつ化合物はデータ集合 S^- に分けておく．このデータ分割は薬物動態学の専門家がクリアランス経路予測を行う際に有効だと考えたものである．ここで分割された各データセットにおけるクリアランス経路の内訳を表 1 に示す．このデータ分割により，クリアランス経路の候補がそれぞれほぼ 3 種類にまで絞られていることが分かる．

経路名	3A4	2C9	2D6	Renal	OATP	Total
データ集合 S^+	52	1	18	23	0	94
データ集合 S^-	0	11	0	18	18	47
Total	52	12	18	41	18	141

表 1 各データセットにおける，クリアランス経路ごとのデータ内訳
Table 1 The number of data for each clearance pathway in datasetS^+ and datasetS^-

4. Boosting Algorithm

ブースティング（Boosting)[6] とは，教師あり学習を実行するための機械学習メタアルゴリズムの一種であり，「一連の弱学習器（Weak Learner) をまとめることで強い学習器を生成できるか？」という Kearns の疑問[8] に基づき考案された手法である．

Boosting は自由度の高い手法であり，例題ごとの分布に従い弱学習器（弱判別器）を繰り返し学習させ，それらを強い学習器の一部とするというものである．一般的には，得られた弱学習器にその正確さに応じた重み付けを行い，その結果から例題ごとの分布の見直しを行う．すなわち，誤判別されたデータは重みを増やし，正しく判別できたデータは重みを減らす．この一連の動作により，次の弱学習器は今までにうまく判別できなかった例題を重点的に判別するようになる．最終的に生成した弱判別器同士で重みを加味した多数決を行い判別を行う．

4.1 AdaBoost

AdaBoost は 1999 年に Freud と Schapire によって提案されたアルゴリズムであり[9]，Boosting のアルゴリズムで最も有名なもののひとつである．AdaBoost は一般的に，他学習法に比べて過学習が起こりにくいとされているが，例題の分布の更新に指数関数が使われており，例題に外れ値やノイズが存在する場合，それらの重みが指数関数的に増大し，外れ値やノイズに反応する弱判別器が生成されるという欠点があることが知られている．

Vol.2009-BIO-17 No.10
2009/5/25

4.2 MadaBoost

MadaBoost は 2000 年に Domingo，Watanabe らによって，提案された Boosting のアルゴリズムであり，AdaBoost を改良したものになっている[10]．MadaBoost は，例題の分布を更新を行う際に，誤判別された例は重みを変えずに，正しく判別できた例のみ重みを減らす．これにより外れ値の重みが極端に大きくなることなく，判別が行うことができる．

5. 実　　験

5.1 クリアランス経路予測システムの構築

図 1 は本研究で構築した予測システムのモデル構成である．このモデルは複数のクリアランスパスを持つデータセットを判別するために，**one-versus-the-rest 法**を用いている．one-versus-the-rest 法は多クラス問題を二値判別器で解く手法であり，$\{C_1, C_2, \cdots, C_n\}$ の各クラスそれぞれについて，クラス C_i を正例とし，残りを負例とした二値判別を全てのクラスについて行うものである．本実験では，各クリアランス経路をクラスとし，5 つの二値判別器を用いた予測システムを用いる．

このシステムから得られる解には通常の “単一解”に加え，複数の判別器から正例と判別された場合の “複数解”，どの判別器からも正例と判別されない “解なし”が考えられるが，本研究では複数解，解なしを認めることにする．これは，一般的に薬物には複数のクリアランス経路をもつ可能性があるものが多く，薬学の見地から見れば複数解という予測も十分妥当であると判断されるためである．

5.2 判　別　器

判別には Boosting を用いる．Boosting のアルゴリズムには，AdaBoost，MadaBoost を使い，弱判別器クラスとして不等式型，挟み込み型と名付けた 2 種類を使用した．

5.2.1 不 等 式 型

以下のような弱判別器を要素にもつ弱判別器クラスを “不等式型”と定義する．

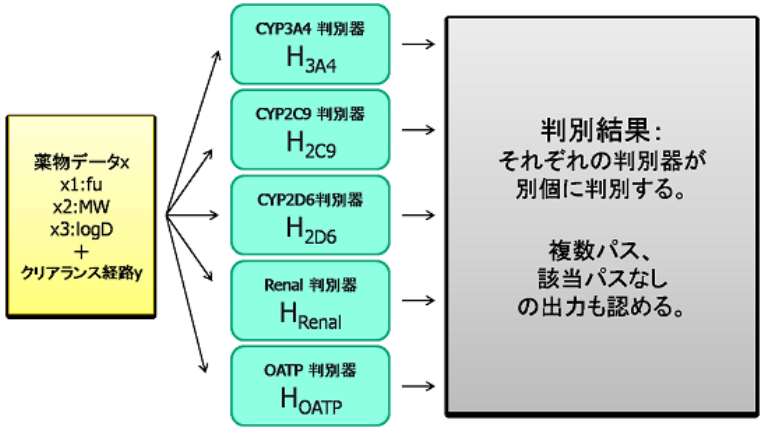

図 1　クリアランス経路予測システム
Fig. 1　Drug clearance pathway prediction system

S^+は正例クラスである例題を含む集合とする．

$i \in \{\mathrm{fu}, \mathrm{MW}, \mathrm{logD}\}$ に対し，

$$h_{leq}(\boldsymbol{x}) = \begin{cases} +1 & \{x_i \leq x_i^+ \mid \boldsymbol{x}^+ \in S^+\} \\ -1 & Otherwise \end{cases} \tag{1}$$

or

$$h_{geq}(\boldsymbol{x}) = \begin{cases} +1 & \{x_i \geq x_i^+ \mid \boldsymbol{x}^+ \in S^+\} \\ -1 & Otherwise \end{cases} \tag{2}$$

これは，正例クラスに属する例題の特徴量をひとつ選び，その特徴量よりも大きい（または小さい）ものをもつ例題を正例と判別する弱判別器クラスである．

5.2.2 挟み込み型

以下のような弱判別器を要素にもつ弱判別器クラスを “挟み込み型”と定義する．

S^+は正例クラスである例題を含む集合とする．
$i \in \{\mathrm{fu}, \mathrm{MW}, \mathrm{logD}\}$ に対し，

$$h(\boldsymbol{x}) = \begin{cases} +1 & \{x_{1i}^+ \leq x_i \leq x_{2i}^+ \mid \boldsymbol{x}_1^+, \boldsymbol{x}_2^+ \in S^+\} \; and \; ※ \\ -1 & Otherwise \end{cases} \tag{3}$$

※ x_{1i}, x_{2i}の間に S^+の要素を 5 つ以上含む．

これは，正例クラスに属する例題を 2 つ選び出し，特徴量をひとつ選び，それぞれの特徴量の間に値をもつ例題を正例と判別する弱判別器クラスである．

5.3 実験 A：構成方法の比較

上記した 2 種類のアルゴリズムと弱判別器クラスを組み合わせて以下の 4 種類の判別器を作成する．

- $H_{mada}^{\leq}$:弱仮説クラスに不等式型，アルゴリズムに MadaBoost を使用した判別器
- $H_{ada}^{\leq}$:弱仮説クラスに不等式型，アルゴリズムに AdaBoost を使用した判別器
- $H_{mada}^{||}$:弱仮説クラスに挟み込み型，アルゴリズムに MadaBoost を使用した判別器
- $H_{ada}^{||}$:弱仮説クラスに挟み込み型，アルゴリズムに AdaBoost を使用した判別器

上記の各判別器において，要素となる弱判別器の個数は 1〜40 個とし，このなかで最も精度が良くなる弱判別器の個数を探索する．
精度測定には**適合率（overall precision)** を用いた．

$$\text{overall precision} = \frac{\text{正しく予測できたデータの総数}}{\text{単一解と予測されたデータの総数}} \tag{4}$$

ここで Leave-one-out 法を用いて精度を評価し，最も汎化性能の良い判別器を求める．

5.4 実験 B：判別領域の他手法との比較

実験 A の結果から最も精度の良い判別器を選択する．この判別器で全データを学習データとして用いた予測を行い，その判別領域を求める．そして，判別領域の解釈性を矩形領域法から得られた判別領域のそれと比較，考察する．

6. 先行研究の手法による計算

6.1 サポートベクターマシンによる予測実験

サポートベクターマシン（SVM) は，1990 年代に V.N.Vapnik らによって考案された二値分類を行う機械学習の手法で，分類する二つのクラス間のマージンを最大化するような判別超平面を作成する．カーネルトリックを用いることで，**計算時間を抑えかつ複雑な曲面の判別面を作成することができる**ため，機械学習の手法として広く用いられている[11]．

年本ら[3)4)] は SVM を用いて，前節で示した one-veusus-the-rest 法を用いたクリアランス経路予測システムを提案した．SVM のカーネルには Gaussian Kearnel を使い，ソフトウェアには SVM^{light} を用いた．また，誤差の指標には F 値（F-measure) を用いた．F 値とは，再現率（recall) と適合率（precision) の調和平均の値で，下式のように表わされる．

$$\mathrm{F} = 2 \cdot \frac{\mathrm{recall} \cdot \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}} \tag{5}$$

各 SVM はソフトマージンとカーネルに対するパラメータ C, σ を持ち，これらを調整することで F 値が最大になる予測を行う SVM の組み合わせを求めることができた．

筆者らは以上の条件で，3 節で説明した最新のデータ集合 S^+ およびデータ集合 S^- に対して SVM を用いた予測実験を行った．なお，各 SVM の予測精度を評価するために，今回は交差確認法（cross validation) の一種である Leave-one-out 法を用いた．

予測実験の結果，4 つの特徴量をもつデータからクリアランス経路を適合率が 82.3% で，37 分 54 秒という高速な計算時間で予測できた[⋆1]．

6.2 矩形領域法による予測実験

矩形領域法（Rectangular method)[4)5)] は，当課題のために年本，草間らが開発した手法で，人間の目から分かりやすく理解できることに重きを置いた二値判別の手法である．

各特徴量ごとに上限，下限（境界）をそれぞれ設定し，全ての条件を満たすようなデータを正例だと判別する．得られる判別領域は（超）直方体となり，その内部と外部で二値判別を行う．この領域の境界のとり方には自由度があるが，判別の F 値が最大となり，かつそのうちで矩形の体積が最小となるような矩形領域を全探索によって探し出す．

以上がこの手法の大まかな流れである．この手法は次元数の増加に伴い，計算量が大きくなるという欠点があるが，判別領域の形が矩形になること，境界が各特徴量で独立であることなどから，**判別が視覚的に理解しやすく，最小限の情報で領域の定義が伝達できるので，専門家間での知識共有が容易である**といえる．

今回，筆者らは矩形領域法を用いた予測を新たに 3 節で説明したデータ集合 S^+ および

⋆1 CPU に Opteron 2.4GHz を使用した．

Vol.2009-BIO-17 No.10
2009/5/25

データ集合 S^- に対して行った．はじめに，5.1 節で説明した one-versus-the-rest 法を用いた構成を使用し，SVM による実験と同様に精度の評価するために Leave-one-out 法を用いる．次に，同様の実験を全データを学習データとして用いて行い，この実験により得られた矩形領域の解釈性を考察する．

予測実験の結果，適合率は 69.1% ，計算時間は 3 時間 56 分 10 秒と SVM による予測には劣るが，生成された矩形領域は人の目からも視覚的に判断しやすく，解釈性の高い結果が得られた[⋆1]．

7. 実験結果

7.1 実験 A：Boosting における各判別器の精度比較

図 2 はそれぞれのデータセットに対して，判別実験を行った結果である．各設定に対して，青色の軸がデータ集合 S^+ の精度，赤色の軸がデータ集合 S^- の精度，緑の点が計算時間をそれぞれ表している．図から不等式型の弱判別器クラスよりも挟み込み型の弱判別器クラスを使用した判別器のほうが計算時間はかかるが，精度が高いことが確認できる．

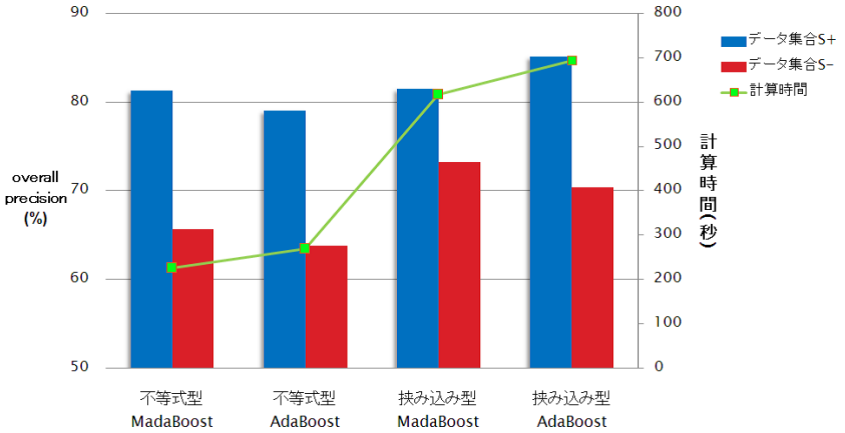

図 2 Leave-one-out 法を用いた精度評価
Fig. 2 Overall precision and calculation time

次に，最も高い精度をもった判別器，すなわちデータ集合 S^+ を挟み込み型の AdaBoost で判別したものと，データ集合 S^- を挟み込み型の MadaBoost で判別したものについて，判別器が予測したクリアランス経路と真のクリアランス経路との関係をデータ数でまとめたものを表 2，表 3 に示す．表 4 はデータ集合 S^+ と S^- による判別結果を合計したものである．これらの表は各データに対する予測結果の詳細である．横方向に表を見ると，各クリアランス経路を解にもつデータがどのように予測されたかが載っており，縦方向に表を見ると，各判別器が検出したデータのクリアランス経路ごとの数の内訳がある．結果として，それぞれのデータセットにおける最良の判別器を組み合わせ，予測器全体の適合率が 81.4% という高精度を実現できた．

$H_{ada}^{\mid\mid}$ データ集合 S^+	predicted pathway								
	3A4	2C9	2D6	Renal	OATP	複数解	解なし	Total	recall
3A4（解）	41	0	2	0	0	1	8	52	79%
2C9（解）	1	0	0	0	0	0	0	1	0%
2D6（解）	2	0	3	2	0	6	5	18	17%
Renal（解）	3	0	0	13	0	1	6	23	57%
OATP（解）	0	0	0	0	0	0	0	0	-
Total	47	0	5	15	0	8	19	94	
precision	87%	-	60%	87%	-				
overall precision	**85.1%**								

表 2 Leave-one-out 法を用いた Boosting の予測結果 (データ集合 S^+)
Table 2 Performance of Boosting algorithm with Leave-one-out method(datasetS^+)

7.2 他手法との精度比較

前節で記した SVM，矩形領域法で行った実験と，本実験の精度および実行時間を比較したものをそれぞれ図 3，図 4 に示す．Boosting による予測は，精度に関しては，矩形領域法よりも高く，おおよそ SVM に匹敵していることが分かる．また，実行時間に関しても，矩形領域法よりも圧倒的に速く，SVM を使った実行時間よりもわずかに速いことが分かる．

7.3 実験 B：矩形領域法との判別領域比較

まず，Boosting で得られた判別領域をデータ集合 S^+ について図 5 に，データ集合 S^- について図 6 に示す[⋆1]．Boosting から得られた領域は矩形領域と似通っているが，評価基

⋆1 CPU に Opteron 2.4GHz を使用した．

⋆1 使用した判別器は実験 A で最も精度が高かったものである．すなわち

Vol.2009-BIO-17 No.10
2009/5/25

$H^{\vdash\vdash}_{mada}$ データ集合 S^-	predicted pathway								
	3A4	2C9	2D6	Renal	OATP	複数解	解なし	Total	recall
3A4（解）	0	0	0	0	0	0	0	0	-
2C9（解）	0	3	0	0	3	1	4	11	27%
2D6（解）	0	0	0	0	0	0	0	0	-
Renal（解）	0	0	0	11	1	0	6	18	61%
OATP（解）	0	2	0	2	8	2	4	18	44%
Total	0	5	0	13	12	3	14	47	
precision	-	60%	-	85%	67%				
overall precision	**73.3%**								

表 3 Leave-one-out 法を用いた Boosting の予測結果 (データ集合 S^-)
Table 3 Performance of Boosting algorithm with Leave-one-out method(datasetS^-)

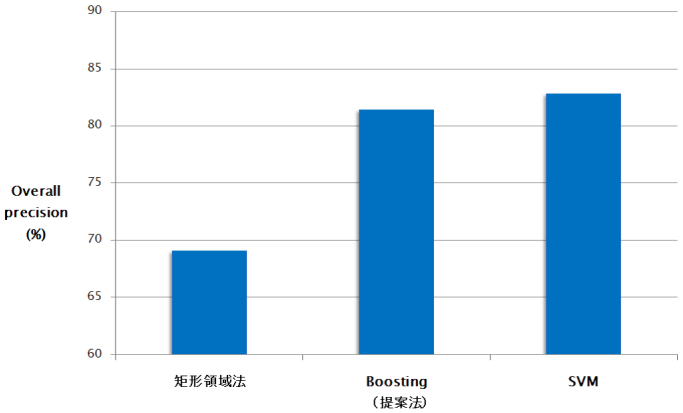

図 3 他手法との判別実験比較 (精度)
Fig. 3 Overall precision(Rectangular method, Boosting, SVM)

Total	predicted pathway								
	3A4	2C9	2D6	Renal	OATP	複数解	解なし	Total	recall
3A4（解）	41	0	2	0	0	1	8	52	79%
2C9（解）	1	3	0	0	3	1	4	12	25%
2D6（解）	2	0	3	2	0	6	5	18	17%
Renal（解）	3	0	0	24	1	1	12	41	59%
OATP（解）	0	2	0	2	8	2	4	18	44%
Total	47	5	5	28	12	11	33	141	
precision	87%	60%	60%	86%	67%				
overall precision	**81.4%**								

表 4 Leave-one-out 法を用いた Boosting の予測結果 (Total)
Table 4 Performance of Boosting algorithm with Leave-one-out method(Total)

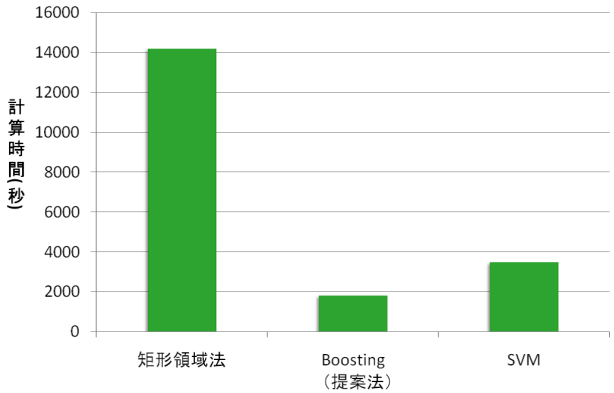

図 4 他手法との判別実験比較 (計算時間)
Fig. 4 Calculation time(Rectangular method, Boosting, SVM)

Vol.2009-BIO-17 No.10
2009/5/25

準や領域決定の自由度の違いから直方体を調整したような形をとる．

表 5 に複数解と解なしの個数を示す．Boosting と矩形領域法を比較した結果，最も顕著な差は複数解の個数であった．これは Boosting が領域の微調整が可能であること示しており，矩形法のとき，矩形の表面近くで誤判別されていた例題を Boosting の場合は矩形を凹ませることで回避することができる．また矩形だと，負例が含まれないように矩形の少し外にある正例をあきらめることがあったが，Boosting の場合は，その正例と負例が近い位置になければ，正例のある部分に膨らませることで正例のみを領域内に収めている．以上の性質から，Boosting では複数解が減り，Leave-one-out 法による検査の範囲内では，正しい判別が増えた．また，外れ値は解なしに入っている．

次に，データ集合 S^+ について，Boosting の場合は Renal が部分的に 2 つに分かれることで両者の共通部分を減らしている．これにより 2D6 が解であるものが複数解に判別される，または Renal に誤判別されることを防いでいる．また，2D6 の領域が拡大し，新たに 3 つの正解を加えることができている．

データ集合 S^- について，図 6 の CYP2C9 の領域に細い帯状のものができ，領域が無理やり正解を含もうとしていることが分かる．CYP2C9 と OATP の判別領域について過学習が起こっていると考えられる．

Method	単一解	複数解	解なし	Total
Boosting	132	2	7	141
Rectangular	105	31	5	141

表 5 Boosting と矩形領域法の判別結果の内訳
Table 5 The Number of unique, multiple, and no solution cases

また，図 7 にデータ集合 S^- におけるクリアランス経路ごとの判別領域を比較したものを示す．ただし，図の分かりやすさを考慮し，MW，logD についての 2 次元にプロットした．赤い丸印が各クリアランス経路の正例，×が負例である．オレンジの枠で囲まれた部分が矩形領域法の解であり，青で塗りつぶされた部分が Boosting で得られた判別領域である．

8. 結　　論

本研究では，Boosting のアルゴリズムを用いて，精度良く，かつ視覚的に分かりやすい

データ集合 S^+ に対し，$H^{||}_{ada}$，データ集合 S^- に対し，$H^{||}_{mada}$ を使用した．

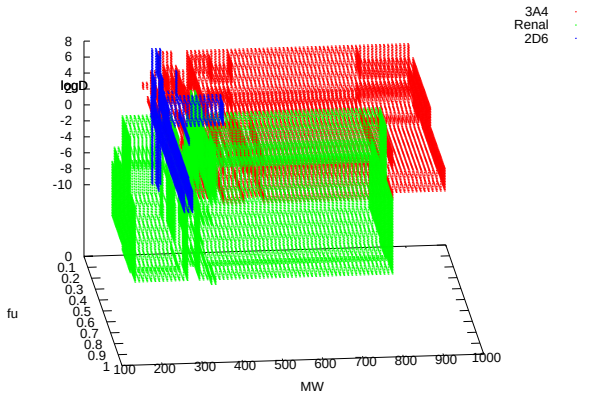

図 5 データ集合 S^+ における各予測器の判別領域
Fig. 5 Regions obtained for CYP3A4, Renal and CYP2D6(datasetS^+)

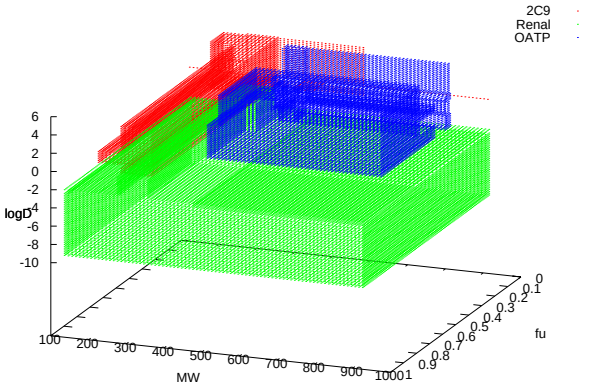

図 6 データ集合 S^- における各予測器の判別領域
Fig. 6 Regions obtained for CYP2C9, Renal and OATP(datasetS^-)

Vol.2009-BIO-17 No.10
2009/5/25

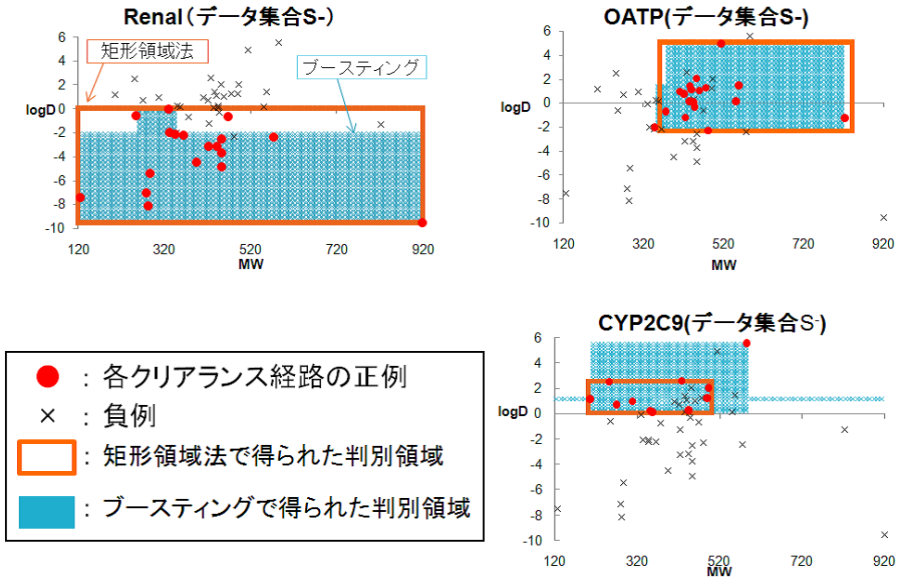

図 7 矩形領域法との判別領域比較
Fig. 7 Comparison of obtained regions between Boosting and Rectangular method

判別領域をもつような判別器を作成する薬物クリアランス経路予測システムを構築した．また，矩形領域法，サポートベクターマシンとの比較を行い，矩形領域法ではうまく判別しきれなかった部分を，矩形を調整，または２つの矩形を用いるような判別領域を作成することで，複数解や誤判別を回避することができた．また，サポートベクターマシンと汎化誤差を比較した結果，ほぼ同等の精度を得ることができた．しかし，今回提案した 2 種類の弱判別器クラスから作成される判別領域は自由度が高くなり，非線形境界を持つサポートベクターマシンよりは解釈しやすいが，矩形領域法と比べれば少し分かりにくいものとなった．今後は弱判別器クラスの制限などを行い，解釈性を向上させることが望まれる．

8.1 今後の課題

本実験では，Boosting を用いて，SVM に匹敵する精度をもった予測を行うことができたが，判別領域の解釈性の高さでは矩形領域法に劣っていた．今後は，解釈性を向上させるために本実験から得られた判別領域に後処理を施し，精度をあまり落とさずに矩形に近く，解釈性がより高い判別領域に整形する工夫を行っていきたい．前述したとおり，現状の矩形領域法は計算量の観点から高次元データに対応できないという欠点があったが，本実験で構成した判別器に判別領域を矩形領域に整形する機構を組み込むことで，これを高次元の例題に対しても高速に動作する矩形領域法の近似解法とすることを試みたい．

謝辞 本研究の一部は，(財）大川情報通信基金 2008 年度研究助成の支援を受けて実施された．

参 考 文 献

1) 厚生労働省医薬品産業実態調査, 2005.
2) 杉山 雄一, 楠原洋之編:“分子薬物動態学”,pp.2-28,pp99-153, 南山堂, 日本, 2008.
3) 年本 広太, 草間 真紀子, 前田 和哉, 杉山 雄一, 秋山 泰: “機械学習を用いた薬物のクリアランス経路予測”, 情報処理学会研究報告, 2008-BIO-13, pp.43-48, 2008.
4) Kouta Toshimoto, Makiko Kusama, Kazuya Maeda, Yuichi Sugiyama, and Yutaka Akiyama: “In silico prediction of major drug clearance pathways by machine learning techniques”, 23rd Annual Meeting of the Japanese Society fot the Study of Xenobiotics, Kumamoto Japan, October 2008.
5) Makiko Kusama, Kouta Toshimoto, Kazuya Maeda, Yuka Hirai, Satoki Imai, Koji Chiba, Yutaka Akiyama, and Yuichi Sugiyama: “Classification of major clearance pathways of drugs based on physicochemical parameters”, 23rd Annual Meeting of the Japanese Society fot the Study of Xenobiotics, Kumamoto Japan, October 2008.
6) Y. Freund, Boosting a weak learning algorithm by majority, Information and Computation, Vol. 121, no. 2, pp.256-285, 1995
7) 大村 恒雄, 石村 巽, 藤井 義明:“P450 の分子生物学”,pp.1-13, 講談社, 2003.
8) Michael Kearns:“Thoughts on Hypothesis Boosting.”, Unpublished manuscript, 1988.
9) Y. Freund and R. Schapire:“A short introduction to boosting”, Journal of Japanese Society for Artificial Intelligence, Vol. 14, no. 5, pp.771-780, 1999.
10) Carlos Domingo and Osamu Watanabe:“MadaBoost: A Modification of AdaBoost” Proceedings of the Thirteenth Annual Conference on Computational Learning Theory:pp.180-189, 2000.
11) B. E. Boser, I. M. Guyon, and V. N. Vapnik: “A training algorithm for optimal margin classifiers”, in 5th Annual ACM Workshop on COLT, D. Haussler, ed., pp.144-152, ACM Press, 1992.