Vol.2010-BIO-21 No.20
2010/6/19

ウェブアプリケーションによる薬物クリアランス経路予測

堀田 駿[†1] 年本広太[†1] 池田和史[†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知ることは薬物動態を解析する上で重要な情報となる．そこで，先行研究として化合物の 4 つの物理化学的特徴量（fup，MW，LogD，charge）からサポートベクターマシン（SVM），矩形領域法，Boosting の 3 種類の機械学習の手法を用いてクリアランス経路を予測する研究が各々おこなわれてきた．本研究では，これらの 3 手法による予測結果をいくつかの方法により統合し，単独の予測システムよりも高精度の予測の実現を目指した．また，この予測システムを創薬現場の専門家が実際に使用できるウェブアプリケーションとして開発を行った．

Drug clearance pathway prediction on a web application

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

In the pharmacokinetics study of drugs, it is very important to know major clearance pathway of the drug in human body. We have developed, in previous studies, three methods to predict major clearance pathway from a few physicochemical parameters (fup, MW, LogD, charge) of the drug, using machine learning techniques: support vector machine (SVM), rectangular method, and boosting technique, respectively. In this study, we integrated the prediction results from the three techniques, aiming to build more accurate prediction system than using an individual method. Also, we have developed a web application for the integrated prediction system so that drug development experts can easily and widely utilize our software.

1. はじめに

新薬の開発は基礎研究，非臨床研究，臨床研究を経て承認申請される．そして，現在，基礎研究に 2〜3 年，非臨床研究に 3〜5 年，臨床研究に 3〜7 年の年月と数十から数百億ドルの費用が必要になり，必要な費用は年々増加している．[1] しかも，実際に上市される確率はわずか 0.01%程しかない．[2] このように，新薬の開発には多くの時間と費用が必要となり，開発の後期や上市された後問題が見つかると大きな損失となる．この問題の解決策の 1 つとして薬物の体内動態の解析がある．

薬物の体内動態は大きく分けると吸収（Absorption），分布（Distribution），代謝（Metabolism），排泄（Excretion）の 4 つに分けられその頭文字をとって ADME と呼ばれる．体内動態を解析するための重要な情報にクリアランス経路がある．クリアランス経路とは，臓器が薬物を代謝・排泄する経路のことである．クリアランス経路を知ることができれば，不適切な化合物を特定できる可能性がある．しかし，クリアランス経路を知るためには臨床試験が必要であり，新薬の開発の早期に特定することはできない．

このクリアランス経路を早期に特定するための方法として計算機上でのクリアランス経路予測がある．先行研究として，化合物の基本的な 4 つの物理化学的特徴量からその化合物が主要な 5 つのクリアランス経路のどの経路で代謝・排泄されるかをサポートベクターマシン（SVM）[3)4)]，矩形領域法[6)7)8)]，Boosting[9)] の 3 種類の機械学習の手法を用いて予測を行うシステムの開発がされて来た．そこで，本研究では，これらの 3 手法による予測結果をいくつかの方法により統合させ，単独の予測法よりも高精度の予測を目指す．さらに，もっとも精度のよくなった統合法を使用し予測システムを作成し，創薬現場の専門家が実際に使用できるようにウェブアプリケーションとして実装した．

2. 薬物データ

機械学習の学習データとして共著者である東京大学の杉山雄一教授らのグループが収集した 140 個の化合物データを用いた．

†1 東京工業大学 大学院情報理工学研究科
Graduate School of Information Science and Engineering, 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

2.1 入力パラメータ

機械学習の入力としては，4 つの基本的な物理化学的特徴量である血漿中タンパク質非結合律（fup），分子量（MW），分配係数（$\log D$），電荷（charge）を使用した．化合物は電荷によって正の電荷をもつ（anion），負の電荷をもつ（cation），電荷をもたない（neutral），正と負の両方の部分をもつ（zwitter）の 4 種類に大別できるが，この研究においては zwitter の化合物は極めて少量しか存在しなかったため除外し，また cation と neutral の化合物を一つにまとめて取り扱った．

2.2 クリアランス経路

厳密には薬物の代謝・排泄を行っている部位は体中のいたるところに存在するが，本研究ではクリアランス経路を肝臓と腎臓について考え，3 つのカテゴリに属する 5 種類のクリアランス経路（Renal，CYP3A4，CYP2C9，CYP2D6，OATP）に絞って実験を行った．

2.2.1 腎排泄（Renal）

腎臓では血液の塩類濃度の調節，老廃物の排泄，水の排泄による尿の生成，薬物の未変化体，代謝物の排泄など生体の維持に重要な機能を担っている．[10] 本研究では，薬物の未変化体が代謝されずに排泄される場合を考えた．一般に親水性の高い薬物は腎排泄されやすく，疎水性の高い薬物は肝代謝されやすいことが知られている．

2.2.2 Cytochrome P450 による肝臓内代謝

Cytochrome P450（以下 CYP）は水酸化酵素ファミリーの総称であり，その重要な役割は体内の薬物の不活性化，あるいは排泄しやすい化合物に変換させる（＝薬物代謝）ことであり，薬物代謝反応の 8 割に関与するとも言われている．現在までに様々な種類の CYP が見つかっており，動物ではその大部分が肝臓に存在する．本研究では，クリアランス経路として，CYP のなかでも薬物代謝に関与が大きい 3 つの酵素（CYP3A4, CYP2C9, CYP2D6）に大きく分類することにした．

CYP3A4 は全肝 CYP 中最大の約 30 %を占める CYP の中心的存在であり，カルシウム拮抗薬や抗生物質などの代謝に関与する．また，CYP2C8, CYP2C9, CYP2C18, CYP2C19 は相同性が 80 %以上とよく似ており，その総量は肝で約 20 %を占める．その中で最も含量が多い CYP2C9 を代表として用いた．CYP2C9 は血糖降下薬や抗凝固薬などの代謝に関与する．CYP2D6 は全肝 CYP の約 2 %ほどにすぎないが向精神薬や循環器用剤，呼吸器用剤などの多くの医薬品の代謝に関与する重要な酵素である．[11]

2.2.3 トランスポータータンパク質による取り込み

トランスポータータンパク質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であり，血中から細胞内への物質の取り込みや逆に細胞内から排出などを行う．

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薬物を肝臓に移行させる代表的なファミリーの一つである有機アニオントランスポーター（Organic anion transporting polypeptide 以下 OATP）ファミリーをクリアランス経路の対象とした．

2.3 薬物データの分布

全 140 のデータを charge によってわけると，表 1 のように，anion の化合物は Renal，CYP2C9，OATP のいずれか，cation or neutral の化合物は Renal，CYP3A4，CYP2D6 のいずれかのクリアランス経路であると判断できる．

表 1 各クリアランス経路における全 140 データの分布
Table 1 The distribution of all drug data in each clearance pathway

経路名	Renal	CYP3A4	CYP2C9	CYP2D6	OATP	Total
anion	18	0	11	0	18	47
cation or neutral	23	52	0	18	0	93
Total	41	52	11	18	18	140

3. 薬物クリアランス経路予測手法

3.1 SVM を用いた経路予測[3)4)]

SVM[12] は教師あり学習を用いるパターン識別手法の一つである．SVM の目的は高次元特徴空間でうまく分離する超平面を学習することである．SVM は境界面から最も近いデータ（サポートベクトル）との距離（マージン）が最大になるように n 次元特徴空間を線形分離する．また，線形分離が困難なことがほとんどであるため，学習データを非線形の写像を使用し別の特徴空間に変換してから線形分離を行う（カーネルトリック）．

年本ら[3)4)] はこの SVM を用いた経路予測において，カーネルに Gaussian Kernel を使用し，SVM のプログラムとして SVM^{light}[13] を使用した．そして，SVM は二値判別プログラムなので，各クリアランス経路ごとにその経路であるものを正例，それ以外を負例として学習を行い，各クリアランス経路それぞれにおいて入力データが属するかどうかを判別した．ただし，この方法では正例と負例の割合が極端に偏ってしまうため，ランダムオーバーサンプリング[5] を行い正例と負例を 200 個:200 個として学習をさせた．ランダムオーバー

Vol.2010-BIO-21 No.20
2010/6/19

サンプリングとは少数のデータに対して重複を許してランダムに選出することで疑似的にデータを増やす方法である．

本研究では，統合を行う際のスコアとして，SVM を使用した予測では，入力ベクトルと分離超平面との距離を用いた．

3.2 矩形領域法を用いた経路予測[6)7)8)]

矩形領域法（Rectangular method）はクリアランス経路予測のために年本，草間ら[6)7)8)]の開発した独自の手法であり，人間が判別境界を視覚的に容易に理解できることを重視している．基本的なアルゴリズムとしては，以下のとおりである．

Rectangular method

（1） データを多次元空間にプロットする．
（2） 各辺が座標軸と平行な矩形を探索し，その矩形の内部を正例，外部を負例と判断して recall と precision を測定し f 値を求める．
（3） 探索された超矩形の中から f 値が最大でかつ体積が最小の矩形を判別領域とする．
（4） 判別矩形の内部を正例，外部を負例として以降の判別を行う．

f 値とは，再現率（recall）と適合率（precision）の調和平均のことであり，以下のように計算を行う．

$$\begin{aligned} \mathrm{f} &= 2 \cdot \frac{\mathrm{recall} \cdot \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}} \\ &= \frac{2TP}{FN + FP + 2TP} \end{aligned} \tag{1}$$

年本，草間らは学習データを charge によって {anion} と {cation or neutral} に分けた後，矩形領域法を使用し入力データが矩形の内部か外部かを調べ判別を行った．charge によって分けたのは，矩形領域法に使用する特徴量は連続した値であることが望ましく，また，charge によってわけることで，クリアランス経路の候補を 5 つから 3 つに絞ることができるためである．ただし，この予測方法では各 charge にデータが存在しないため予測を行えない経路が存在する．

本研究では，統合を行う際のスコアとして，矩形領域法では，判別矩形の中心のスコアを 1，矩形上を 0，判別矩形と中心が同じで各辺の長さが 2 倍の矩形上からその外部すべてをスコア-1 として，1 から-1 まで線形に値が変化するように設定した．(図 1)

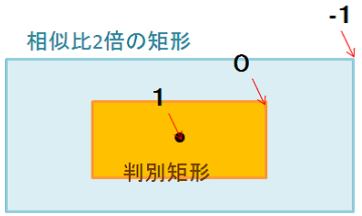

図 1 矩形領域法のスコア
Fig. 1 The score of rectangular method

3.3 Boosting を用いた経路予測[9)]

Boosting[14)] は単純だがあまり性能の良くない学習アルゴリズム（弱学習器）を元に，性能のよい学習アルゴリズムを構成する方法である．基本的なアルゴリズムとしては，以下のとおりである．

Boosting Algorithm

（1） 弱仮説クラス H から分布 P に対して最も良い弱判別器 h_{opt} を選択する
（2） h_{opt} の重要度 α_{opt} を算出し，それまでの重要度に足し合わせる
（3） 統合決定関数 $f_{\boldsymbol{\lambda}_t}$ を $f_{\boldsymbol{\lambda}_t}(\boldsymbol{x}) := \sum_{h \in H} \alpha_h h(\boldsymbol{x})$ とする
（4） 分布 P を更新する
（5） 終了条件を満たすまで 1 から 4 を繰り返す
（6） 統合決定関数 $f_{\boldsymbol{\lambda}_t}$ を出力する

実際の学習では，総合決定関数の最適化と同義である損失関数の最小化を行う．

池田ら[9)] は，弱学習器に挟み込み型という正例に属するデータを 2 つ選び出し，特徴量を一つ選び，それぞれの特徴量の間に値を持つデータを正例と判別する弱学習を用い，損失関数としては，charge が {cation or neutral} のデータについては Adaboost を，{anion} の化合物については，Madaboost を使用した．そして，矩形領域法と同様に charge によってデータを分けた後，Boosting アルゴリズムにより学習させ，入力データの判別を行った．この手法も矩形領域法と同様にデータが存在しないため予測を行えない経路が存在する．

3.3.1 Adaboost[15)]

Adaboost は最も有名な Boosting アルゴリズムであり，損失関数に $U_{ada}(z) = e^z$ を使用

Vol.2010-BIO-21 No.20
2010/6/19

する．Adaboost は他学習法に比べ，過学習が起こりにくいとされている．しかし，z の値が大きくなると損失も際限なく大きくなってしまうため外れ値やノイズに弱いことが知られている．

3.3.2 Madaboost[16)]

Madaboost は Adaboost にフィルタリングの機能を付けて，ノイズに強くすることを目的として作られた手法であり，損失関数には，

$$U_{mada}(z) = \begin{cases} \frac{1}{2}e^{2z} & z \leq 0 \\ z + \frac{1}{2} & z > 0 \end{cases}$$

を用いる．

これは，予測が外れたものに対して，指数関数でなく一次関数を用いるため，外れ値の影響が比較的小さくなり，判別機が外れ値の影響を過度に受けないように努めるようになる．

本研究では，統合を行う際のスコアとして，Boosting を使用した予測では，総合決定関数の値をスコアとした．

4. 予測結果の統合方法

3 種類のクリアランス経路予測の手法の統合方法として，それぞれの手法による予測結果を多数決により統合する方法と，SVM を用いて統合する方法を行った．SVM のプログラムとしては SVM^{light} を使用した．

4.1 多数決による統合

多数決による統合では SVM，矩形領域法，Boosting の 3 手法の内 2 手法以上で正例だと判断されたものを正例とし，それ以外を負例とした．また，矩形領域法・Boosting において anion の時の CYP3A4 などモデルがなく判別できない経路については SVM による予測のみにより正例と負例を判別した．

4.2 SVM による統合

SVM による統合では，学習用のデータの構造として，各クリアランス経路における 3 手法の予測のスコアを 3 次元データとして用いる方法と，5 種類すべてのクリアランス経路における 3 手法の予測のスコアを（3 × 5 =）15 次元データとして用いる方法を試みた．また，矩形領域法・Boosting において anion の時の CYP3A4 などモデルがなく判別できない経路についてはスコアを 0 として学習させる方法と，矩形領域法では-1，Boosting ではすべての弱学習器で負例と判別された時の値を使用して学習させる方法の 2 通りを試みた．

表 2 統合方法
Table 2 The ways for integration

	統合①	統合②	統合③	統合④	統合⑤
統合方法	多数決	SVM			
判別できない経路のスコア		矩形領域法: 0 ブースティング: 0		矩形領域法: -1 ブースティング: すべての弱学習器で負例のときの値	
入力		3 次元データ	15 次元データ	3 次元データ	15 次元データ

5. 実　　験

5.1 実験 1: f 値が最大になる SVM^{light} のオプションの探索

10-fold cross validation を使用し，SVM^{light} で設定できる正と負のサンプルの誤差の割合（以下 J）を { デフォルト値，正:負= 1:2 になる値，正:負= 1:1 になる値 } の 3 通りと，同じく設定できる誤識別率とマージンの比重（以下 C）を { デフォルト値，0.001，0.01，0.1，1，10，100} の 7 通りの計 21 通りの値の組み合わせを使用し，表 2 の統合②から統合⑤について f 値が最大になる J，C の組み合わせを探索した．

5.2 実験 2: 全 140 の化合物データにおける各統合方法の予測性能の比較

表 2 の 5 種類の統合方法について予測性能を比較した．SVM を使用する統合では，J，C の値は実験 1 で f 値を最大にした組み合わせとし，また 10-fold cross validation を使用して予測性能を比較した．

5.3 実験 3: 新規の 35 の化合物データにおける各統合方法の予測性能の比較

表 2 に示す 5 種類の統合方法について，元の 140 の化合物データとは異なる新規の 35 の化合物データにおける予測性能を比較した．この 35 の化合物データは 2004 年以降に日米欧で承認された医薬品の中から主なクリアランス経路が Renal，CYP3A4，CYP2C9，CYP2D6，OATP であるデータを抽出しており，先に述べた 140 の化合物データと同じく東京大学薬学部の杉山雄一教授らのグループから頂いた．SVM を使用する統合では，J，C の値は実験 1 で f 値を最大にした組み合わせを用いた．新規の 35 の化合物データの分布は以下のとおりである．

表 3 各クリアランス経路における新規 35 データの分布
Table 3 The distribution of new drug data in each clearance pathway

経路名	Renal	CYP3A4	CYP2C9	CYP2D6	OATP	Total
anion	3	1	1	0	0	5
cation or neutral	12	18	0	0	0	30
Total	15	19	1	0	0	35

6. 実験結果

6.1 実験 1: f 値が最大になる SVM^{light} のオプションの探索

表 4 は統合②から⑤において，各 J の値に対して，C の値を変化させ各クリアランス経路において最大となった f 値の和を示した．

この結果をみると統合②から⑤のいずれも正例と負例の比が 1:2 の時の f 値の合計が最大であったので，以下の実験では，J には正例と負例の比を 1:2 にする値を使用し，C にはこの J の値に対し，各クリアランス経路で最大の f 値を出した値を使用する．(表 5)

表 4 各統合法の J の値による f 値の変化
Table 4 the value of f-measure according to value of J in each integrated methods

	J=default 値	J（正例：負例 = 1:2）	J （正例：負例 = 1:1）
統合②	2.749	3.385	3.271
統合③	3.122	3.323	3.171
統合④	2.857	3.440	3.265
統合⑤	3.213	3.382	3.215

表 5 各クリアランス経路において最大の f 値を出した C の値（J（正例：負例 = 1:2））
Table 5 The value of C which puts out the maximum f-measure in each clearance pathway（J（positive : negative = 1 : 2））

	Renal	CYP3A4	CYP2C9	CYP2D6	OATP
統合②	1	10	0.01	0.01	0.1
統合③	0.01	default 値	0.01	0.1	1
統合④	0.1	0.01	0.01	1	default 値
統合⑤	0.01	10	0.001	0.1	10

6.2 実験 2: 全 140 の化合物データにおける各統合方法の予測性能の比較

まず矩形領域法，SVM，ブースティングの 3 手法を単独で使用した場合における 10-fold cross validation をした際の Recall（Total），Precision（Total），f 値を表 6 に示す．

そして，実験 2（学習用 140 データにおける各統合方法の予測性能の比較）の結果の Recall（Total），Precision（Total），f 値をまとめると表 7 のようになる．

この表 7 から，f 値を比べると統合②が最も良く，次に統合④が良いという結果が分かる．つまり，SVM による統合の入力に 3 次元データを与えたほうが，15 次元データを与えたものと多数決による統合よりも若干良いと判断できる．

表 6 3 手法の予測性能
Table 6 Predictive performance of three techniques

	矩形領域法	SVM	ブースティング
Recall（Total）	0.63	0.86	0.49
Precision（Total）	0.61	0.64	0.73
f 値	0.62	0.73	0.59

表 7 実験 2（学習用 140 データにおける各統合方法の予測性能の比較）の結果
Table 7 Comparison result of predictive performances of each integrated methods

	統合①	統合②	統合③	統合④	統合⑤
（Total） Recall	0.69	0.82	0.81	0.86	0.83
（Total） Precision	0.70	0.65	0.60	0.62	0.59
f 値	0.70	0.73	0.69	0.72	0.69

6.3 実験 3: 新規の 35 の化合物データにおける各統合方法の予測性能の比較

まず矩形領域法，SVM，ブースティングの 3 手法における新規の 35 のデータを使用し予測した時の Recall（Total），Precision（Total），f 値を表 8 に示す．

そして，実験 3（新規 35 データにおける各統合方法の予測性能の比較）の Total の Recall, Precision, f 値をまとめると表 9 のようになる

この表 9 から，新規 35 データの場合は統合①の多数決による統合が最も f 値がよいという結果が分かる．

Vol.2010-BIO-21 No.20
2010/6/19

表 8 3 手法の予測性能（新規 35 データ）
Table 8 Predictive performance of three techniques (new drug data)

	矩形領域法	SVM	ブースティング
Recall（Total）	0.66	0.86	0.66
Precision（Total）	0.88	0.68	0.79
f 値	0.75	0.76	0.72

表 9 実験 3（新規 35 データにおける各統合方法の予測性能の比較）の結果
Table 9 Comparison result of predictive performances of each integrated methods (new drug data)

	統合①	統合②	統合③	統合④	統合⑤
（Total） Recall	0.77	0.77	0.60	0.69	0.71
（Total） Precision	0.77	0.64	0.49	0.47	0.43
f 値	0.77	0.70	0.54	0.56	0.54

6.4 考　　察

実験 2（学習用 140 データにおける各統合方法の予測性能の比較）においては統合①の多数決による統合の f 値は 3 番目であったが，実験 3（新規 35 データにおける各統合方法の予測性能の比較）においては最も f 値が良くなった．そのため，統合①は外部データに強いのではないかと考えウェブアプリケーションへの実装は統合①の多数決によるものを使用することにした．

しかし，今回外部データは 35 と少なく，データの偏りも見られるため偶然外部データの時に多数決による統合の精度が良いという結果になったとも考えられる．また，実験 2 におけるどの統合方法も，矩形領域法とブースティングによる予測よりは予測性能がよいが，SVM による予測の予測性能と比較した場合同じかそれよりも低い予測性能の方法しかなかった．よって，今後新たな統合方法の探索と予測性能の確認を行い更なる予測精度の向上を図る必要がある．

7. ウェブアプリケーションの開発

上記の fup，MW，log D，charge の 4 つの物理化学的特徴量から SVM，矩形領域法，Boosting の 3 手法でクリアランス経路を予測し，さらにその 3 手法による出力を統合させ，最終的なクリアランス経路の予測を行うウェブアプリケーションを開発した．予測を行うための入力方式として，Individual Entry，Batch Entry，File Upload の 3 種類を選択することができる．このウェブアプリケーションは現在認証が必要ではあるが，http://www.bi.cs.titech.ac.jp/PKPD/test/ に公開している．

図 2 入力画面
Fig. 2 Input screen

7.1 ウェブアプリケーションの動作

7.1.1 Individual Entry

この入力方式では，1 つの化合物の特徴量（fup，MW，log D，charge）を各スペースに入力する．（図 2）

fup 血漿タンパク質非結合率（fup）は血中での血漿タンパク質と結合していない割合を表したものであり，この値が大きいほど細胞膜を透過し組織中に拡散しやすい．fup は割合であるので 0 から 1 の実数値を入力する．

MW 分子量（MW）は化合物の 1mol あたりの質量のことである．

log D 混ざらない 2 つ以上の溶媒に化合物を入れた時，その化合物は各溶媒の間で一定の割合で分配される．薬物における分配係数（log D）とは，油としての n-オクタノールと水を使用し，各溶媒の間で分配される薬物濃度の比の常用対数を指し，脂溶性の指標となる．この値が正のとき脂溶性が高く，負のとき低い．

charge 電荷（charge）には anion（負の電荷を持つ化合物），cation（正の電荷を持つ化合物），neutral（電荷を持たない化合物），zwitter（正と負の両方の電荷を持つ化合物）の 4 種類が存在する．この予測システムでは学習したデータの中 zwitter の化合物

は含まれておらず，また予測システムでは charge が cation と neutral の化合物を区別せず一つのグループとして予測を行ったので，入力としても anion と cation or neutral のどちらかを選択する．

化合物の物理化学的特徴量を入力した後，“ Predict ”ボタンをクリックすることで予測結果のページが出力される（図 3）．

この予測結果の見方は，Input information には自分の入力した特徴量が，Prediction results には上から順にクリアランス経路名，3 手法による予測の出力の統合結果，矩形領域法による予測結果，ブースティングによる予測結果，SVM による予測結果が出力される．予測の統合結果は，3 種類の予測手法すべてで正例と予測した時“ ＋＋ ”，2 種類の予測手法で正例または no data と SVM でのみ正例と予測した時“ ＋ ”，それ以外の時“ － ”で表している．矩形領域法，ブースティング，SVM の予測結果はそれぞれのスコアを棒グラフで表してあり，棒グラフが中心よりも右側まで達しているなら正例，中心まで達していないなら負例を表している．また，正例のときは赤い棒グラフ，負例のときは青い棒グラフで示している．

予測結果ページの Classification boundaries の“ Rectangular method ”の“ 3-D figure ”ボタン，“ 2-D projection figure ”ボタン，“ Boosting method ”の“ 2-D projection figure ”ボタンをクリックすると，それぞれ矩形領域法の 3 次元グラフにおける判別境界と入力データのプロットしたグラフ，矩形領域法の判別境界と入力データの投影図，ブースティングの判別境界を入力データの位置で切断したグラフが出力される．

7.1.2 Batch Entry

この入力方式では，1 行に 1 つの化合物の特徴量を（化合物名），(fup)，(MW)，(log D)，(charge）の順番でコンマでつないで値を入力する．charge は anion ならば 1，cation or neutral ならば 0 とする．(図 2)

入力をした後“ Predict ”ボタンを押すことで入力したすべての化合物について予測を行う．この Batch Entry で予測を行った場合，図 3 の予測結果ページとは異なり，表の各行に左から入力した化合物名，各クリアランス経路の予測結果が表示され，それが入力された化合物の数だけ縦に連なって表示される（図 4）．

そして，各化合物名の下にある“ detail ”ボタンを押すと，各化合物に対する，図 3 の予測結果ページが出力される．

7.1.3 File Upload

この入力方式では，Batch Entry と同様に 1 行に 1 つの化合物の特徴量を（化合物名），

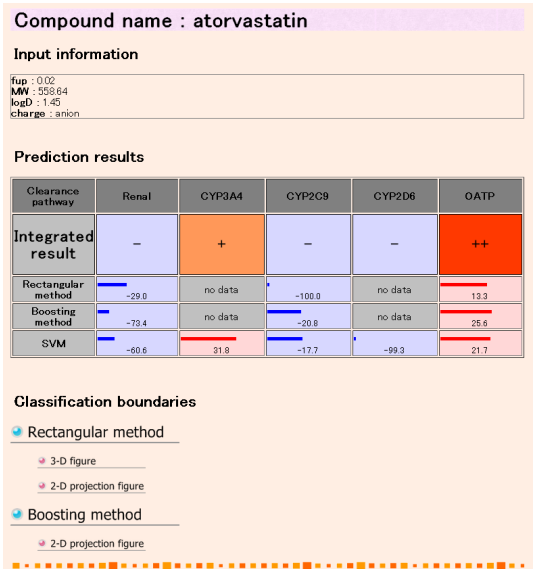

図 3 予測結果画面
Fig. 3 Prediction result screen

（fup），（MW），（log D），（charge）の順番に並べた CSV 形式のファイルを入力とする．charge は anion ならば 1，cation or neutral ならば 0 とする（図 2）

出力の形式としては Batch Entry と同様である．

8. 結　　論

本研究では，矩形領域法，SVM，ブースティングの 3 手法によるクリアランス経路予測の結果を統合し，精度の向上を図った．実験した 5 種類の統合方法の中では，SVM に 3 次元の入力を使う方法が元の 140 のデータについては最も精度がよく，新規の 35 のデータについては多数決による統合が最も精度が良いという結果になった．よって，ウェブアプリケーションではすでに学習したデータではなく新しいデータの予測が望まれるので，多数決による統合の方が新規データでの性能が期待できると考え，選択した．

8.1 今後の課題

今回ウェブアプリケーションに実装した多数決による統合方法は，第七章の実験結果をみ

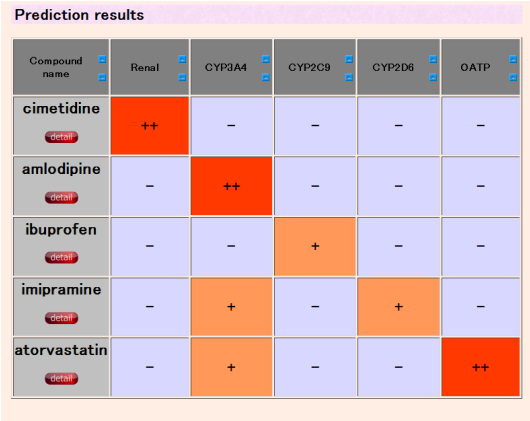

図 4 Batch processing mode での予測結果画面
Fig. 4 Prediction result screen on the batch processing mode

ると外部データに対しては矩形領域法，SVM，ブースティングのどの手法よりも良い精度でクリアランス経路の判別ができていたが，10-fold cross validation の結果では統合をしない SVM の f 値や 3 次元入力を SVM を用いて統合した場合よりも精度が悪かった．そのため，統合方法をさらに吟味する必要がある．また，新規の 35 のデータはかなりの偏りがあり，また量も少ないのでさらなる検証が必要かと考えられる．

また，今回２つ以上のクリアランス経路において正例だと予測されたものについてはそれぞれ予測されたクリアランス経路について１回ずつ数えているが，今回の学習用のデータにないというだけで実際は２つのクリアランス経路をもつものも存在する．そのため，２つ以上のクリアランス経路において正例だと予測されたものの扱いを吟味する必要がある．

参 考 文 献

1) 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）
2) 治田 俊志,“ 日本の医薬品開発におけるトランスレーショナルリサーチの役割 ”，DDS 22: 36-42．2007 .
3) 年本 広太，草間 真紀子，前田 和哉，杉山 雄一，秋山 泰.“ 医薬品の物理化学的特性に基づいた薬物動態プロファイリング （II）．”，第 24 回日本 DDS 学会．2008．Jun
4) Kouta Toshimoto，*et al.*:“ *In silico* prediction of major drug clearance pathways by machine learning techniques．”，23rd Annual Meeting of the Japanese Society for the Study of Xenobiotics．2008．Oct
5) Liu Y, An A, Huang X:“ Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles ”, In the Proceedings of the 10 Pacific-Asia Conference on knowledge discovery and data mining（PAKDD ’06）, Singapore, pp.107-118（2006）.
6) 草間 真紀子，平井 由香，前田 和哉，今井 覚己，千葉 康司，年本 広太，秋山 泰，杉山 雄一．:“ 医薬品の物理化学的特性に基づいた薬物動態プロファイリング （I）．”，第 24 回日本 DDS 学会．2008．Jun
7) Makiko Kusama，*et al.*“ Classification of major clearance pathways of drugs based on physicochemical parameters．”，23rd Annual Meeting of the Japanese Society for the Study of Xenobiotics．2008．Oct
8) Makiko Kusama，*et al.*“ In Silico Classification of Major Clearance Pathways of Drugs with their Physiochemical Parameters ”，Drug Metab Dispos （in press）
9) Kazushi Ikeda，*et al.*“ Prediction of drug clearance pathway by boosting algorithm ”，IPSJ SIG Technical Report，2009-BIO-17（10）:1-8，June，2009.
10) 西垣隆一郎，堀江利治，伊藤智夫:“ 薬物動態学”，pp.49-52 丸善，東京（1998）
11) 加藤隆一，鎌滝哲也:“ 薬物代謝学－医療薬学・毒性学の基礎として－ ”，pp.9-61, 東京化学同人，東京（1995）.
12) Nello Cristianini:“ サポートベクターマシン入門”，pp.129-149，共立出版，東京（2005）.
13) SVM^{light}, http://svmlight.joachims.org/
14) 金森敬文，畑埜晃平，渡辺治:“ ブースティング－学習アルゴリズムの設計技法－ ”,pp.1-85, 森北出版株式会社，2006.
15) Y.Freund and R. Schapire:“ A short introduction to boosting”, Journal of Japanese Society for Artificial Intelligence，14（5）:771-780，September，1999.
16) C．Domingo and O．Watanabe,“ A modification of AdaBoost ”，Proc．the 13th Annual Conference on Computational Learning Theory，ACM pp.180-189，2000.