Vol.2011-BIO-25 No.7
2011/6/23

能動学習法を利用した薬物クリアランス経路予測の改良

稲吉一馬[†1]　石田貴士[†1]　前田和哉[†2]
杉山雄一[†2]　秋山泰[†1]

我々は薬物のクリアランス経路を機械学習を用いて予測するシステムを構築してきた．予測性能は実用的にも有用なレベルには到達しているものの，予測精度は飽和状態にある．これは学習データが少ないことが原因であり，さらなる予測精度の向上には新たなデータの追加採取が望ましい．しかし新しい薬物に対するヒトのクリアランス経路の決定のコストが高いため，学習データの追加は容易には行えない．そこで本研究では「どの新規薬物に対してクリアランス経路の調査実験を行えば，より予測システムの精度が向上するか」を示唆するために，能動学習の手法によって実験すべきデータの優先度を推定する方法を提案する．

Improvement of the clearance pathway prediction by using active learning

Kazuma Inayoshi,[†1] Takashi Ishida,[†1]
Kazuya Maeda,[†2] Yuichi Sugiyama[†2]
and Yutaka Akiyama[†1]

We have been developing prediction systems of the clearance pathway of drugs. The performance of the system is insufficient because of the shortage of the training data, and the additional training data is required for the improvement of the system. However, *in vivo* experiments of human clearance pathway for new compounds require huge cost and time, and thus it is difficult to gather much training data. In this study, we propose a method which suggests high-priority compounds to be examined by using active learning.

1. はじめに

新薬の開発コストを抑えるために，開発初期での薬物スクリーニングが近年重要になってきている[1)]．この要請を応えるために，我々は薬物の体内クリアランス経路の予測を行ってきた．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．我々は化合物の基本的な 4 つの物理化学的特徴量から，機械学習の一手法であるサポートベクターマシン（SVM）を用いて 5 つの経路についてクリアランス経路を予測を行うシステムを構築してきた[2)]．また解釈が明解なモデルとして，多次元での超矩形を用いた『矩形領域法』と称する手法によるシステムも構築した[3)]．

一般的に機械学習では，良い予測性能を達成するためには十分な数の偏りの無い学習データが重要である．しかしながら我々のクリアランス経路予測システムの構築では学習に利用することができたデータが少数であったため，特に一部の経路において予測システムの性能は必ずしも十分とは言えないものであった．データ数を増やすことでこの問題は改善されることが見込まれるが，薬物に対するヒトのクリアランス経路のデータを増やすためには，長い期間や高額な費用を要する[1)]．そのため「予測システムの性能の向上への寄与が大きいと予想される薬物」を能動的に選択し，その薬物のクリアランス経路を明らかにして学習データに追加することで，予測システムの性能の向上が可能となれば，非常に有用である．このように追加学習データを意図的に選択する方法は能動学習法[4)] と呼ばれ，機械学習において 1 つの分野を成している．

能動学習はデータが入出力の組からなる教師付き学習に利用でき，特に，出力データの採取に要するコストが大きいが，出力を観測する前の入力データは多量に用意できるような問題で有用である．我々の扱うクリアランス経路予測問題の状況はこれに合致しており，特に有効であると考える．

創薬の分野においては Warmuth らによって「活性を有する化合物をできるだけ多く，少ない試験で見つけ出す」ことを目的とした追加データ選択の方法が提案され[5)]，成功を収めており，この分野での能動学習の適用は有用であることが期待される．

本研究では「クリアランス経路予測問題に能動学習を適用することで，将来的に予測精度

†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

表 1 各クリアランス経路における全 176 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	合計
データ数	56	71	13	18	18	176

※ 各薬物の主要なクリアランス経路を示しており，排他的にいずれかのクラスに属する

が向上するか」を明らかにすることを目的とし，Warmuth らによって提案された能動学習法を含む 5 つの手法について薬物データを用いて予測精度向上の比較を行った．またその結果より各能動学習法の有用性について論じ，新たな能動学習法のアイデアを提案した．

2. 薬物データ

薬物クリアランス経路予測の学習データとして関連研究で用いられている 176 個の化合物データを用いた[3]．

2.1 入力記述子

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電荷（charge），分子量（MW），分配係数（LogD），血漿タンパク質非結合率（fup）を使用した．化合物は電荷によって負の電荷をもつ（anion），正の電荷をもつ（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，使用する薬物データには zwitter の化合物は含まれていない．また cation と neutral の化合物をまとめ，負の電荷を持つ（1）か否（0）かとして取り扱った．

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，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[6]．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[7]．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．

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．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 176 個の薬物データのうち，各クリアランス経路を持つものがいくつあるかを**表 1** に表した．各薬物データは Renal，CYP3A4，CYP2C9，CYP2D6，OATP のいずれかのクラスに排他的に属する．

3. SVM によるクリアランス経路予測

3.1 SVM（サポートベクターマシン）

SVM（サポートベクターマシン）は教師付き学習に基づく 2 クラス分類器である．SVM は以下の線形モデルを用いて，$y(\mathbf{x})$ の符号により分類を行う．

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b$$

ここで $\phi(\mathbf{x})$ は非線形な特徴空間への写像を行う関数である．本研究中では $y(\mathbf{x}_i) > 0$ となるデータ $\mathbf{x}_i$ を正例と予測，$y(\mathbf{x}_i) < 0$ となるデータ $\mathbf{x}_i$ を負例と予測されるとする．SVM は**図 1** のように，分離境界と学習データの間の最短距離マージンを最大にするような $(\mathbf{w}, b)$ を選択する．また SVM は線形の識別器であるが，カーネル関数の利用により非線形分類器の学習を行うことが可能である．

本研究においては学習ステップを繰り返し行い，暫定の予測システムが次の学習データ選択に与える影響が大きいため，学習の精度は重要である．よって分類器として SVM を用いることとする．実装にはカーネル関数に Gaussian Kernel を使用し，SVM のプログラムには SVM^{Light}[8] を使用した．

3.2 予測システムの評価

構築された予測システムは，データを正例，もしくは負例と予測する．このとき予測された各データは**表 2** のように TP，TN，FP，FN のいずれかとなる．本研究では f 値を予測

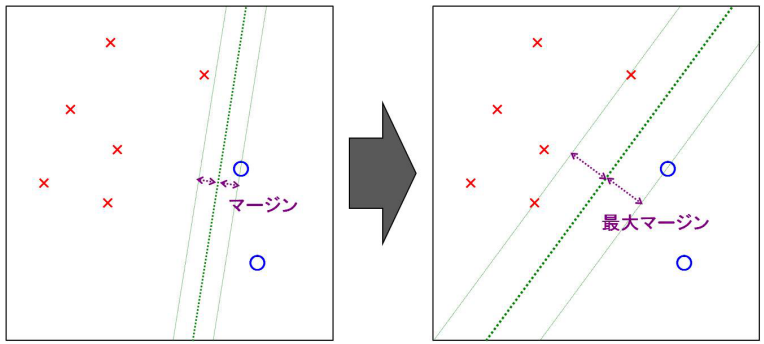

図 1 マージンの最大化
Fig. 1 Maximization of a margin

精度の評価として用いる．f 値は再現率と適合率の調和平均として計算される．

- 再現率（recall）：実際の正例の中での正例と予測されたデータの割合．
  $\mathrm{recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$
- 適合率（precision）：正例と予測されたデータの中での実際の正例の割合．
  $\mathrm{precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$
- f 値（f-measure）：再現率と適合率の調和平均．

$$\mathrm{f} = 2 \cdot \frac{\mathrm{recall} \cdot \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}$$

4. 能動学習法

能動学習の目的は，予測器構築のための学習データを集める段階で予測器の精度向上への寄与が大きいと予想されるデータを選ぶことである．学習データが多いほど一般に学習の精度は良くなるが，実世界ではデータの採取コストが高く，大量の学習データを採取することが困難であることが多い．このような限られたデータ数で効率の良い学習を行う際に能動学習は特に有用となる．

表 2 データの真のラベルと予測ラベルの関係
Table 2 Relations between the true label and the predicted label

	正例と予測	負例と予測
正例	TP（真陽性）	FN（偽陰性）
負例	FP（偽陽性）	TN（真陰性）

能動学習の手法は大きく「一括能動学習」と「逐次的能動学習」に分けられる[9]．「一括能動学習」はデータの分布などから一括で採取する学習データを最適化する手法であり，「最適実験計画」とも呼ばれる．一方「逐次的能動学習」は事前にいくらかの学習データを採取し，そのデータにより構築された分離境界に基づく手法である．「一括能動学習」のメリットは事前にデータの採取を行う必要がないため，全ての学習データを最適化することができる．しかしながら適用できる問題が線形回帰問題などの特別なケースに限られ，分類問題に対する適用は困難である．よって本研究では分離境界の推定と学習データの最適化を逐次的に行う「逐次的能動学習」に焦点を絞る．

4.1 探索と搾取

逐次的な学習において，探索搾取問題と呼ばれる問題がある[10]．この問題は「新しい可能性の探索」を行うことで大局的によりよいものを目指す方策と，「古い確証からの搾取」によって現在よいとされるものをより細かく調べる方策とがトレードオフの関係にあることを指している．

分類問題における逐次的能動学習においての「探索」は分離境界を大雑把に定めるために，現在の分離境界に捕われずに広範囲からデータを選ぶことを指し，「搾取」は分離境界をきめ細かく定めるために，分離境界の周りからデータを選ぶことを指す．この問題においては一般的に，学習初期においては探索を優先し，ある程度暫定分離境界が安定してきたら徐々に搾取の割合を増やしていくのが良いと言われている．

4.2 先 行 研 究

逐次的能動学習の枠組みで Warmuth らによって，多量の化合物群の中から目標分子に結合する（＝活性がある）化合物を少ない生化学試験でできるだけ多く見つけ出すデータ選択法が提案されている[5]．この先行研究においては効率的な化合物スクリーニングを目的としており，その目的を達成するために「正例データを多く収集すること」と「分類器の学習精度を向上させること」の 2 点に主眼をおいた，2 つのデータ選択法が提案されている．

我々のクリアランス経路予測問題においては「予測システムの学習精度向上」のみを目的としているが，先行研究で提案される方法は一般の能動学習の枠組みにおける探索と搾取に相当するような，汎用性のある考え方に基づく手法である．そこで今まで能動学習の利用がなされて来なかったクリアランス経路予測問題に対して先行研究における能動学習法を適用し，有用性を調べた．以下ではまず能動学習中で行われる 2 つのデータ選択の戦略について説明し，その後全体の能動学習アルゴリズムの説明を行う．

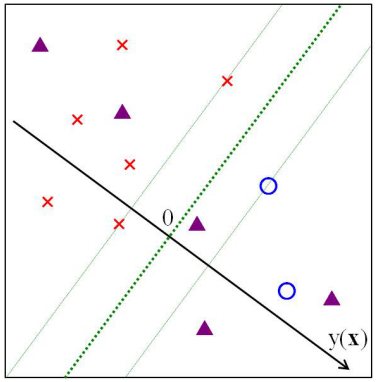

図 2 追加データの選び方
Fig. 2 Selection methods of additional data

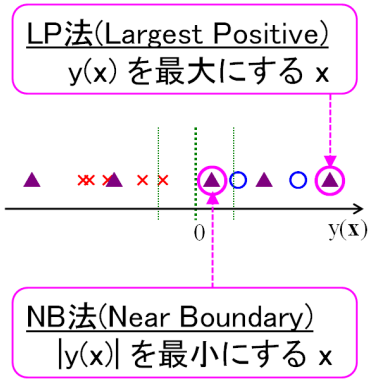

図 3 分離境界からの距離
Fig. 3 Distances from the boundary

4.3 能動的なデータ選択法

図 2 は既に所持している学習データ（○が正例，×が負例）とそのデータより学習された暫定的な分離境界（太字破線），そして学習データとして追加する候補となるデータ（▲）を表している．**図 3** は図 2 に対して，分離境界からの距離を表したものであり，右側にあるものほど $y(\mathbf{x})$ が大きい．以下で 2 つのデータ選択法がそれぞれどの▲を選択するかを，アイデアと共に説明する．

4.3.1 Largest Positive 法（LP 法）

Warmuth らの提案する方法であり，現在のモデルにおいて，最も正例らしいと思われる▲（図 3 では最も右側にある▲）を追加学習データとして選ぶ．SVM によって学習されたモデルから得られるスコアを利用し，具体的には以下の式に基づいてデータ $\mathbf{x_{LP}}$ を選択する．

$$\mathbf{x_{LP}} = \underset{\mathbf{x}}{\operatorname{argmax}}\{y(\mathbf{x})\}$$

我々の扱うクリアランス経路予測の問題では，正例（○）が少数であるクリアランス経路が多いため，少数データをより確実に増やせる LP 法は予測精度の向上させることが見込める．

4.3.2 Nearest Boundary 法（NB 法）

Warmuth らの提案する方法であり，現在のモデルにおいて，最も境界に近い▲を追加学習データとして選ぶ．SVM によって学習されたモデルから得られるスコアを利用し，以下の式に基づいてデータ $\mathbf{x_{NB}}$ を選択する．

$$\mathbf{x_{NB}} = \underset{\mathbf{x}}{\operatorname{argmin}}\{|y(\mathbf{x})|\}$$

この手法は予測結果がより不確かであると思われる現在の分離境界付近のデータを調べることで，少ないデータの追加でより精度の良い分離境界の構築を目指す方法である．

4.4 その他の選択法

上記の 2 つの能動学習法の性能を比較するために，以下の 3 つの選択法を比較手法として用意した．

4.4.1 ランダム選択

現在の分離境界を考慮せず，全候補から一様な確率で選ぶ．この方法は過去の学習データを利用しないため，本質的には一括で学習データを選ぶことと同じである．

4.4.2 Largest Negative 法（LN 法）

LP 法が正例を優先的に選択するのに対し，LN 法は負例を優先的に選択する方法として以下のようにデータ $\mathbf{x_{LN}}$ を選ぶ．

$$\mathbf{x_{LN}} = \underset{\mathbf{x}}{\operatorname{argmax}}\{-y(\mathbf{x})\}$$

4.4.3 Furthest Boundary 法（FB 法）

NB 法が正例を優先的に選択するのに対し，FB 法は現在の分離境界から，以下のように距離が最も遠いデータ $\mathbf{x_{FB}}$ を選ぶ．

$$\mathbf{x_{FB}} = \underset{\mathbf{x}}{\operatorname{argmax}}\{|y(\mathbf{x})|\}$$

4.5 逐次的能動学習アルゴリズム

逐次的能動学習アルゴリズムは，初期状態で n_0 個の学習データを所持しているときに，合計 m 個のデータを新たに採取できる場合において以下に示す流れでより予測精度を向上させるようなデータの選び方を提案する．

(1) 初期学習データ n_0 個を用いて，SVM によって分離境界面を学習する．
(2) データ選択法に基づき n_b 個のデータを選び，ラベルを観測して学習データに加える．
(3) 更新された $n_i (= n_{i-1} + n_\mathrm{b})$ 個の学習データを用いて，SVM によって分離境界面を更新する．
(4) (2)〜(3) を追加データが m 個を超えない範囲で繰り返す．

5. 5つのクリアランス経路予測問題に対する実験

5.1 データセットの説明

5つのクリアランス経路予測問題に対しての能動学習法の有用性を検証するために，4節の5つの能動学習法を適用し，精度向上の比較を行った．学習データには2節で紹介した176個の薬物データを用いた．本実験で用いたデータは全て排他的にいずれかの主要クリアランス経路を持つ問題ではあるが，クリアランス経路を2つ以上持つ薬物を含む問題へも対応するために各クリアランス経路に対して，薬物がそのクリアランス経路を持つか否かの2クラス分類を行う．そのため薬物データは1対他クラス分類器を構築するための5種のデータセットとして別々に扱う．

5.2 能動学習法の性能の比較

本実験ではまず用意した各データセットに対して，全データを用いた場合で最もf値が高くなるSVMの学習パラメータC，γの探索を行った．Cの候補として{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 20, 50, 100}の10通り，γの候補として{0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 20, 50}の10通りを用意し，これを10-fold交差確認法によってf値が最大となるC，γの組み合わせを探索した．

このパラメータを用いて，ランダム選択，NB法，LP法，FB法，LN法の5つの能動学習法が選ぶデータが，それぞれどのようにf値の向上に貢献するかを比較した．実験は**図4**の手順で行った．また逐次能動学習アルゴリズムのパラメータには$n_0 = 20$，$n_\mathrm{b} = 5$を用いた．

実験の手順

(1) 全データを10組に分け(*1)，1組をテスト標本(*2)，残りを訓練標本候補として使用する．
(2) 訓練標本候補より，初期訓練標本としてランダム(*3)にn_0個を選出する．
(3) 現在の訓練標本を用いて，SVMによって予測器を学習する．
(4) 学習された予測器によって，テスト標本を予測しf値を記録する．
(5) 標本選択法に基づき訓練標本候補からn_b個のデータを選び，訓練標本に加える．
(6) 訓練標本候補が無くなるまで(3)〜(5)を繰り返す．

上記の手順を試行1回分として，

- (*1) データの割り方を10通り
- (*2) テスト標本の選び方を10通り
- (*3) 初期標本の選び方を10通り

を試し，計1,000回の試行の平均を実験結果として出力する．

図4 実験の手順
Fig. 4 Protcol of the experiment

5.3 実験結果

表3は各クリアランス経路に対して，10-fold交差確認法により全176データで学習及び評価をした際のf値である．能動学習の実験においてのSVMの学習パラメータには最大のf値となった際のC，γを用いた．

図5は横軸に学習データ数，縦軸にf値を取ったグラフである．各値は1000回の試行の平均値となっている．また**表4**は，全データを用いて学習したときのf値の90%の値を出すために必要な最小学習データ数を表したものである．ここではランダム選択，NB法，LP法の3手法による結果を併記し，ランダム選択よりも性能の良かったケースを太字で示した．

5.4 考察

図5および表4の結果より，全体のf値が0.6〜0.7とある程度良いCYP2C9やOATPではNP法とLP法はともにランダム選択よりもf値の立ち上がりが早く，学習データ数が35〜50個において全データから得られたf値の90%の値を達成している．

一方全体のf値が0.8前後と良く，正例データを多く有するRenalやCYP3A4に関しては，NB法は30〜35個と少数データで全データから得られたf値の90%の値を達成し，他手法に比べ良い性能を出したが，LP法に関しては他手法よりも悪い結果となった．表1か

表3 5つの各クリアランス経路における最適なC，γでのf値
Table 3 F-measures by optimal parameters in each 5 clearance pathway

経路名	Renal	CYP3A4	CYP2C9	CYP2D6	OATP
f値	0.769	0.832	0.600	0.459	0.631

表4 全データで学習したときのf値の90%の値を達成する最小学習データ数
Table 4 Minimum dataset size to achieve 90% of f-measure by all data

経路名	Renal	CYP3A4	CYP2C9	CYP2D6	OATP
NB	**35**	**30**	**45**	**80**	**35**
LP	65	90	**50**	**85**	**40**
R	45	40	85	115	50

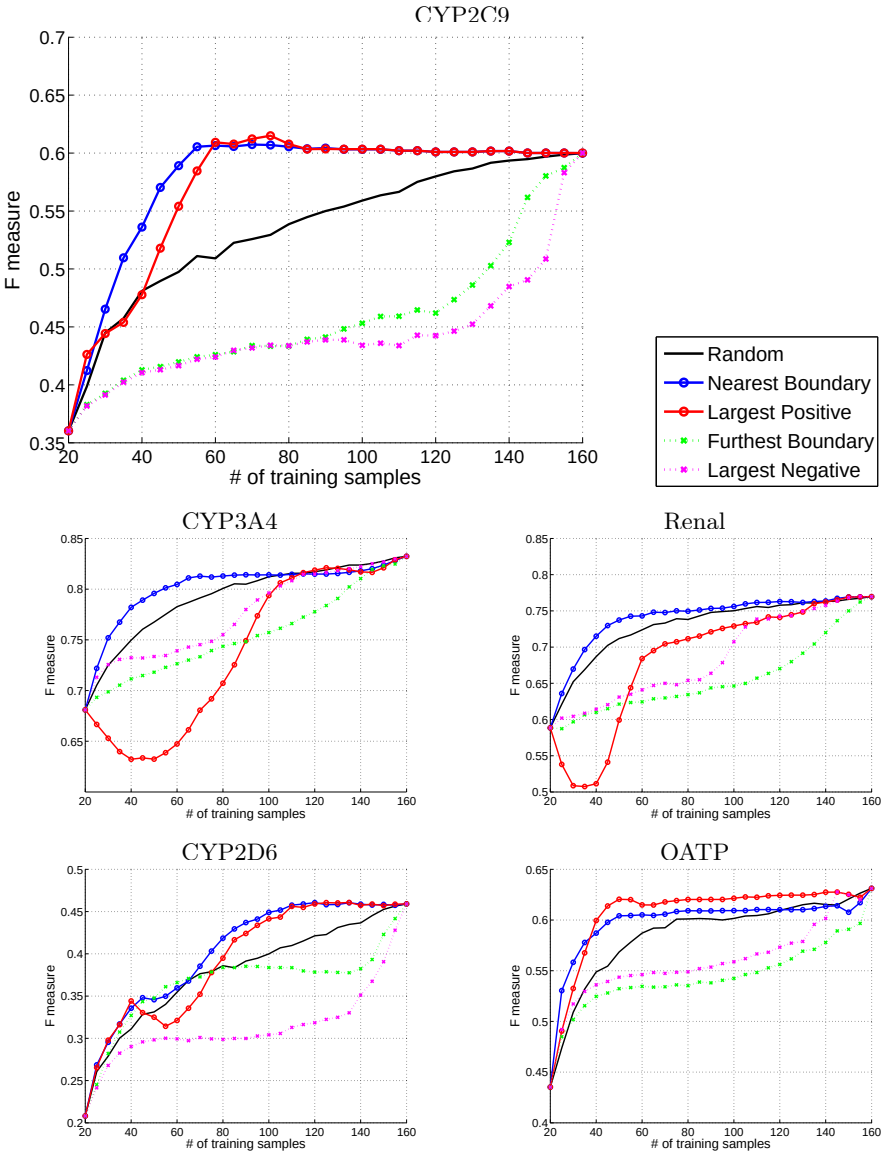

図 5 データ追加による f 値の変化
Fig. 5 Development of f-measure with the increasing data

ら分かるようにこの 2 つのクリアランス経路を持つ薬物は他の経路に比べてその数が多い．f 値の基となる再現率と適合率を個別に確認したところ，再現率の低下，すなわち FN の増加が f 値の低下を引き起こしていることが分かった．これは正例である可能性が高いものから優先的に採取する LP 法が，分離境界から遠いところにある学習データを用いて予測システムを過学習してしまったためと考えられる．

最後に，全データでの学習で達成できた f 値が 0.5 以下であった CYP2D6 では，NB 法と LP 法は 80〜85 個の学習データで 90%の f 値を達成しており，他のクリアランス経路に比べて多くの学習データを要したもののランダム選択よりも早期の f 値向上が達成できた．またこのクリアランス経路が他と異なる点として，学習データ数が 65 個までに着目した場合に FB 法がランダム選択よりも f 値の向上が早いことがグラフから分かる．これは暫定的な分離境界が未熟である段階では，FB 法のように空間の広域に渡りデータ採取を行って分離境界を大まかに決めに行くこと（探索）が有用であるという，能動学習の戦略として広く知られる事実と一致しており，f 値が低い段階においては FB 法はランダム選択よりも f 値向上に貢献できている．以上のことから 5 つのクリアランス経路予測問題においても，予測精度がまだ低い経路の予測については探索に基づく手法が有用であることが示唆された．

6. その他の応用例に対する実験

5 節では正例データ数が多く有する Renal や CYP3A4 において，従前の提案では良い手法の 1 つとして提示されていた LP 法は予測精度向上に不向きであることが確認された．この事実が 5 つのクリアランス経路問題に特有の結果なのか，一般的なデータについても同様に起き得ることなのかを確認するために分類問題のベンチマークとしてよく利用される USPS のデータを用いて追加の実験を行った．

6.1 手書き数字データ USPS[11)]

USPS データは米国郵便公社（U.S. Postal Service）が主催するプロジェクトの一環として，CEDAR に収集された 16 × 16 画素からなる手書き数字データであり，学習問題のベンチマークとしてしばしば用いられる．

- 入力 :各画素の濃度を $-1 \sim 1$ で表したもの．256 次元の入力である．
- クラス :各データは $0, 1, ..., 9$ の 10 クラスのいずれかに属する．
- データ数 :各クラスには 708〜1,553 個のデータが含まれ，全データ数は 9,298 個である．

6.1.1 データセット

USPS データはデータ数が 10,000 個近くあり，この一部をデータセットとして切り出す

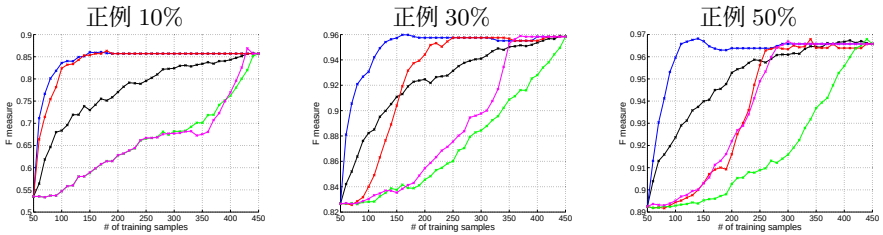

図 6 データ追加による f 値の変化（USPS 問題：クラス“ 8 ”）
Fig. 6 Development of f-measure with the increasing data（USPS: class 8）

ことで実験設定を変化させた．特に正例の比率による LP の性能の差を確認するため，データセットを切り出す際には正例の割合が 10%，30%，50%の 3 つの場合を試すこととした．

本実験では対象としたクラスである（正例）か否（負例）かの 2 クラス分類を行う．実験には {0, 1, ..., 9} のいずれかのクラスより {50 個, 150 個, 250 個 } のデータをランダムに選び，その他のクラスから合計を 500 個にするように残りのデータを選ぶことで，データ数 500 個のデータセットを 30 種を用意した．逐次能動学習アルゴリズムのパラメータには $n_0 = 50$，$n_\mathrm{b} = 10$ を用いた．

6.1.2 実験結果と考察

本実験は“ 0 ”〜“ 9 ”の各数字に対して，正例の比率を 10%，30%，50%とした計 30 個のデータセットにより実験を行ったが，“ 0 ”〜“ 9 ”の各数字が同じような結果を示したため，ここでは“ 8 ”を代表として扱う．**図 6** では，手書き数字データが“ 8 ”か否かを予測する問題についての，正例の比率を 10%，30%，50%と変化させて実験を行った結果である．（線の色は図 5 に準じる）

図 6 より USPS による実験においても，正例の割合の小さなうちは LP 法は NB 法と同様にランダム選択に比べ良い f 値向上を見せているが，正例の割合が増すにつれて LP 法の f 値向上はランダム選択と比べて悪くなっていることが分かる．この結果より，クリアランス経路予測問題についてだけでなく，より一般的な USPS データにおいても LP 法は正例が多く採取できる状況においては予測精度向上への貢献が難しいことが示唆された．

6.2 7 つのクリアランス経路予測（未発表）

現在我々は新たに 2 つのクリアランス経路を追加し，計 7 種のクリアランス経路に対する実験を行っている．の 2 つの新規クリアランス経路（経路 6，経路 7 とする）の詳細については本稿では匿名のままとし，新規データによる実験で得られた結果の一部を公開する．

表 5 7 つの各クリアランス経路の正例数と，最適な C，γ での f 値
Table 5 Number of positive samples and optimal f-measure in each 7 clearance pathway

経路名	Renal	CYP3A4	CYP2C9	CYP2D6	OATP	経路 6	経路 7
正例数	70	82	17	25	12	19	26
f 値	0.722	0.758	0.595	0.313	0.709	0.359	0.338

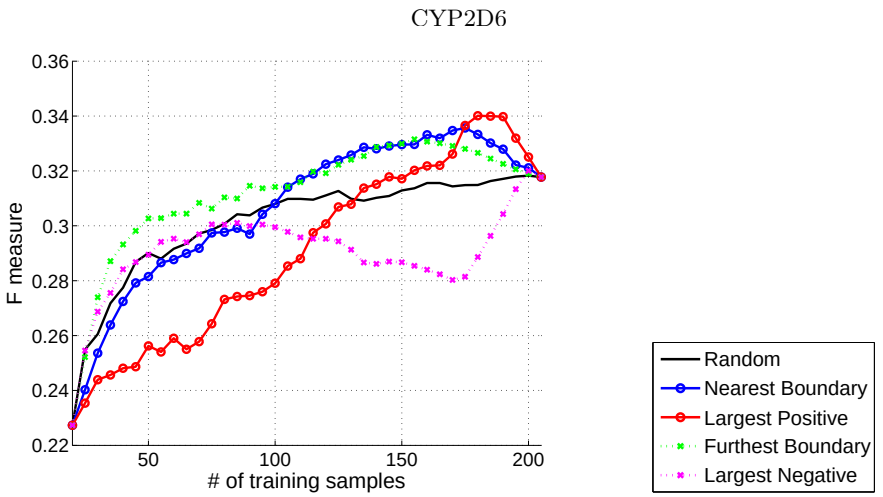

図 7 データ追加による f 値の変化（7 経路問題：CYP2D6）
Fig. 7 Development of f-measure with the increasing data（7 clearance pathway: CYP2D6）

6.2.1 データセット

本実験においてはクリアランス経路を 2 つ以上持つ薬物を含む．そのため薬物データは 1 対他クラス分類器を構築するための別々の 7 種のデータセットとして別々に扱う．

表 5 は 7 つのクリアランス経路をそれぞれ有するデータ数と，全データで学習及び評価して最適化した f 値である．新規データにおいては CYP2D6 と経路 6，経路 7 の 3 種のクリアランス経路に対する f 値が 0.3〜0.4 と低い値にとどまった．

6.2.2 実験結果と考察

図 7 は CYP2D6 に関する学習データ数と f 値の関係を表したグラフである．5 つのクリアランス経路予測の問題においては，図 5 のように CYP2D6 では NB 法は常にランダム選択より高い f 値を出していたが，図 7 においては学習データが 100 個未満の場合において

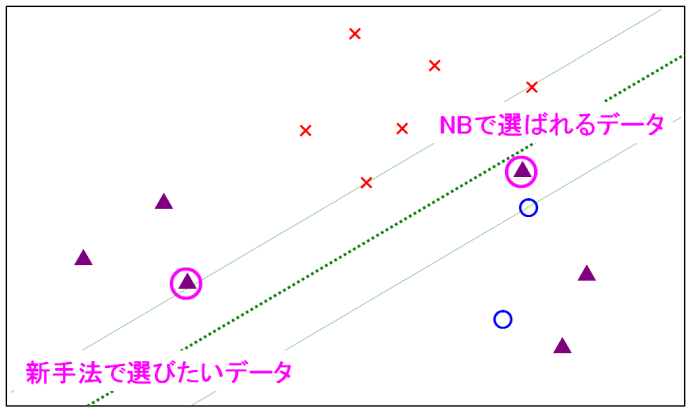

図 8 選びたいデータのイメージ
Fig. 8 Schematic Important of data to be selected

NB 法はランダム選択よりも f 値の向上が悪く，逆に FB 法が最も f 値の向上が早い．このことからも 5.4 節で述べたように分離境界が曖昧な段階では,「搾取」よりも「探索」を優先することが f 値の向上に繋がることが示唆される．

5 節においてランダム選択よりも精度向上に貢献した NB 法が悪くなり，代わりに FB 法での f 値の向上が早かった理由としては，クリアランス経路を増やしたことによって各経路の予測精度が下がったことが考えられる．このように経路の数が増え各経路の予測がより難しくなる場合，探索的な能動学習を優先して行うことが重要になると考えられる．

7. 結　　論

本研究では，SVM によるクリアランス経路予測の精度向上を図るために，能動学習の理論に基づいて効率の良いデータ追加方法を模索した．クリアランス経路予測実験と手書き数字認識の予測実験から，既存手法である NB 法，LP 法が，早期の f 値の向上の目的において有用であるケースとそうでないケースがある場合を確認した．特に LP 法は正例が多く採取できる環境下においては f 値の向上にマイナスに働くケースがあった．また NB 法に関しては暫定 f 値が低い段階では FB 法を選択した方が f 値向上に貢献できる可能性があることが分かった．これは NB 法が分離境界をより細かく決定する「搾取」を行うために，暫定分離境界が不確かな場合は効果的なデータが得られないことに相当すると思われる．またこのような状況で FB 法が f 値向上に貢献しているケースにおいては，FB 法が広範にデータを採取することで，分離境界をおおまかに正しくする「探索」がうまくできているからだと考えられる．

7.1 今後の課題

今回の実験により，既存法として提案されている 2 つの能動学習法が我々の扱う薬物のクリアランス経路予測において常に優位に働くわけではないことが明らかになった．今後は暫定 f 値を推定しながら，f 値が低かった場合に NB 法の代わりとなる方法を採用し，学習された分離境界がある程度安定した段階で NB 法へ切り替える方法を模索することを考えている．また探索と搾取を同時に実現するような，分離境界に近いながらも既存データから距離のあるデータを採取できるような手法についても検討したい．(**図 8**)

参　考　文　献

1) 杉山雄一，楠原洋之：分子薬物動態学，南山堂，pp.2–28，pp.99–153 (2008).
2) 年本広太，草間真紀子，池田和史，堀田駿，前田和哉，杉山雄一，秋山泰：SVM を用いた薬物クリアランス経路予測システムの開発—複数経路予測への拡張と外部データによる評価—，情処研報，Vol.2010-BIO-20，No.8 (2010).
3) Kusama, M., Toshimoto,K., Maeda,K., Hirai,Y., Imai,S., Chiba,K., Akiyama,Y. and Sugiyama,Y.: In Silico Classification of Major Clearance Pathways of Drugs with Their Physiochemical Parameters, *DMD*, Vol.38, pp.1362–1370 (2010).
4) MacKay, D.J.C.: Information-based objective functions for active data selection, *Neural Computation*, Vol.4, No.4, pp.590–604 (1992).
5) Warmuth, M.K., Liao, J., Ratsch, G., Mathieson,M., Putta, S. and Lemmen, C: Active Learning with Support Vector Machines in the Drug Discovery Process, *J.Chem.Inf.Comput.Sci.*, Vol.43, pp.667–673 (2003).
6) 西垣隆一郎，堀江利治，伊藤智夫：薬物動態学，丸善，pp.49–52 (1998).
7) 加藤隆一，鎌滝哲也：薬物代謝学ー医療薬学・毒性学の基礎としてー，東京化学同人，pp.9–61 (1995).
8) *SVM*^{Light}, http://svmlight.joachims.org/
9) 渡辺澄夫，萩原克幸，赤穂昭太郎，本村陽一，福水健次，岡田真人，青柳美輝：学習システムの理論と実現，森北出版，pp.98-131 (2005).
10) March, J.G.: Exploration and exploitation in organizational learning, *Organization science*, Vol.2, No.1, pp.71–87 (1991).
11) Hull, J.J.: A Database for Handwritten Text Recognition Research, *IEEE PAMI*, Vol.16, No.5, pp.550–554 (1994).