修士論文

SVMと学習データの選択を用いた薬効予測システムの構築

同志社大学大学院 生命医科学研究科 生命医科学専攻

博士前期課程 2011年度 115番

宮部　洋太

指導教授　廣安 知之教授

2012年1月25日

Abstract

For the purpose of giving administration criteria of anticancer drug to medical staff, drug effect prediction system is developed in this research. Based on the patient information whose drug effects have already been known, this system presents criteria of drug effect prediction for patients whose drug effects are still unknown. Specifically, the discriminant function in a feature space which determines the effectiveness of the drug is shown by using SVM (Support Vector Machine).

Generally, data points on feature space, which represent patients, are impossible to be divided perfectly from their drug effects. This system realizes limited drug effect predictions by dividing feature space into predictable area and unpredictable area. There is trade-off relationship between the accuracy of predictions and wideness of predictable area. In order to present several criteria having different prediction accuracy and different wideness of predictable area, the method of SVM with multi-objective optimization is proposed. SVM technique is formulated as a multi-objective optimization problem which is demanded not only to minimize the accuracy of training data but also to maximize wideness of predictable area. Adopting SVM technique, validity of several decision criteria can be determined visually, as an additional function.

目 次

1　序論

進行・再発癌に対する治療に一般的に用いられる抗がん剤は効果や副作用に個人差がある．これは体格や年齢，肝臓や腎臓の障害，合併症，過去の治療歴など様々な要因で薬に対する反応が異なるためであり，無効な場合の医療費や副作用が問題となっている．そのため癌患者にとって抗がん剤の効果を事前に予測することは切実な願いである．

近年，バイオマーカが薬の効果と副作用の予測に使われている．例として HER2 の発現量は転移性乳がんの分子標的治療薬であるハーセプチンの投薬判断に用いられており，患者の HER2 の発現量が強陽性の場合，併用することで生存期間が長くなることが証明されている[1)]．画像診断，心電図，及び受容体の量など人体から得られるあらゆる情報が薬の効果や副作用と関係のあるバイオマーカになりうるとされており，投与患者の生体情報と副作用，及び薬効の関係性が見出され，新たなバイオマーカが発見され，薬効予測が可能な抗がん剤の種類が増加することが期待されている．

本研究ではこうしたバイオマーカあるいはバイオマーカ候補の特徴量から SVM(Support Vector Machine)[2)] により薬効を予測し，医療従事者が診断の参考とするシステムの開発を研究目的として掲げている．このシステムはある抗がん剤 A に対する効果が確認されている患者の特徴を基にその効果を分類する基準を学習し，未知の患者の投与前の特徴を入力することで未知の患者に対する投与後の効果を予測するものである．以下にシステムの手順を示す．

step1 患者から予測に有効な特徴 (バイオマーカ or バイオマーカ候補) を抽出

step2 特徴を元に SVM によって判断基準を学習

step3 判断基準を元に効果を予測し，診断の参考にする

step1 において，抽出される学習対象は，一般的に線形分離が不可能な分離不可能問題である場合が多い．分離不可能問題に対する SVM のアプローチは通常，少数の誤りを許容するソフトマージン SVM に適用し，非線形カーネルを用いて特徴空間上の学習ベクトルを高次元空間に写像し，高次元空間上で線形分離する基準を学習することで解決されてきた．しかしながら分布が大きく重複している場合，正確な基準が構築されないことや，意味をなさない複雑な識別線が構築されることが往々にしてあった．

そこで本論文では学習対象の特徴空間を予測可能領域と予測不可能領域に分離する SVM の利用法を提案する．重複領域における学習データの除外を行うことで，片方の class について学習データを完全に分離することが可能となる．これにより分離不可能問題でも学習データに対して限定的に完全な識別が可能とする．また少数の誤識別を許容し，その度合いに応じて完全に分離が可能な範囲を拡張する識別基準を同時に求めることも考慮する．これにより許容される誤識別の度合いに応じてユーザが適切な基準をユーザが選択できることが期待される．さらに本論文では提案手法によって得られた複数の識別基準を表示し，妥当な識別基準を視覚的に検討するためのユーザインタフェースを開発する．

次章ではパタン認識問題と SVM の概要について述べ，3 章で提案手法である SVM の学習法について説明し，多目的最適化アルゴリズム NSGA-II により実装する方法，及びデータセットに対して

行った実験結果について述べる．4 章では提案手法によって得られた結果を視覚的に検討するためのインタフェースの構築について述べる．5 章で結論を述べる．

2　パターン認識

2.1　パターン認識問題

パターン認識[3] は，認識対象がいくつかの概念 (class) に分類できる時，観測されたパターンをそれらの概念のうちのひとつに対応させる処理である．パターン認識では訓練サンプルを計算機に与えて学習させ，その後，テストサンプル（未知データ）が来たときに正しく識別させることを目的とする．

ここで，l 個の観測データ $\{x_i, y_i\}$,$i = 1, \ldots, l$ が与えられているとする．このとき，$x_i \in R^n$ は特徴ベクトルであり, $y_i \in \{-1, 1\}$ はそれぞれの特徴ベクトルに対応する class である．次に，関数 $f : R^n \to R$ が次の式 (2.1), 及び式 (2.2) の条件を満たすものとする．

$$f(x_i) > 0 \text{ if } y_i = 1 \tag{2.1}$$

$$f(x_i) < 0 \text{ if } y_i = -1 \tag{2.2}$$

このような f を識別関数と呼ぶ．識別関数によって，未知のデータ x に対応する class y を

$$y = \mathrm{sgn}(f(x)) \tag{2.3}$$

によって推定することができる．このとき $\mathrm{sgn}(f(x))$ は

$$\mathrm{sgn}(f(x)) = 1 \text{ if } f(x) > 0 \tag{2.4}$$

$$-1 \text{ if } f(x) < 0 \tag{2.5}$$

によって表される符号関数である．

2.2　Support Vector Machine

Support Vector Machine(SVM）は，V. Vapnik などによって提案された，パターン認識の分野において優れた性能を示すことが知られている手法である[2) 4) 5)]．これまでに，数字認識[2)]，テキスト分類[6)]，顔検出[7)] などといった様々なパターン認識に SVM は適用されている．SVM は教師あり学習を用いる識別手法の一つであり，線形 SVM と非線形 SVM に分類される．

ある学習サンプル $x_i \in R^d$ は class$y_i \in \{1, -1\}$ に属し，class 毎に線形分離可能だとすると，その判別関数は重み w を用いて次式で表される．

$$f(x_i) = (w^t x_i) + b \tag{2.6}$$

ここでは b はバイアス項であり，$f(x) = 0$ を満たす点の集合（識別面）が $d-1$ 次元の分類超平面となる．また，この超平面が l 個全ての学習サンプルを分離可能として一意に定まるには制約条件式 (2.7) を満たす必要がある．

$$y_i \cdots ((w^t x_i) + b) \geq 1 \ (i = 1, \ldots, l) \tag{2.7}$$

この時，超平面に最も接近するサンプル（サポートベクトル）と超平面までの距離（マージン）は常に $\frac{1}{\|x\|}$ となり，このマージンを最大化するような w を選ぶことで汎化能力の高い判別関数が推定される．つまり線形 SVM の問題は 式 (2.7) の制約条件の下，$\|x\|^2/2$ を最小化する凸２次計画問題に帰着する．特徴空間が２次元の場合の例を Fig. 1 に示す．ここでは，●が classA の学習データ，○が classB の学習データを表す．

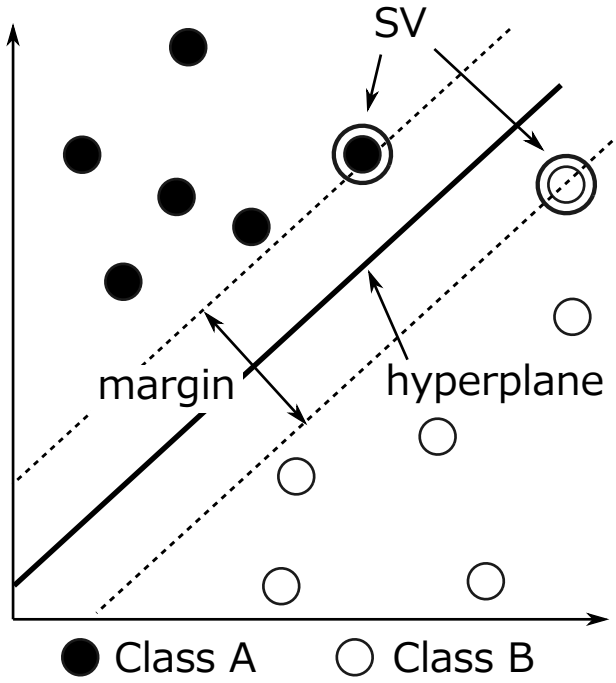


Fig. 2.1　SVM の識別面

一般に，実データに対しては完全な線形分離は困難な場合が多い．そこで若干の誤分類を許容し，その度合いを表す緩和変数 $\xi \geq 0$ と，誤分類とマージン最大化の関係を調節する係数 C を導入することによって，最小化問題は式 (2.8) のように変更される．

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i \\ &\text{s.t. } y_i((w^t x_i) + b) \geq 1 - \xi_i (i = 1, \ldots, l) \end{aligned} \tag{2.8}$$

これは緩和係数の和を小さく，かつ識別能力を高める w を求める問題となる．ここで係数 C を任意に定めることにより，そのトレードオフを決定できる．この最適化問題をを扱いやすい形に変換するためにラングランジュ乗数 $\alpha \geq 0$ を導入すると，最適化における条件として式 (2.9) が導かれる．

$$w = \sum_{i=0}^{l} \alpha_i y_i x_i \tag{2.9}$$

w は学習サンプルの展開式となり，式 (2.10) の相対問題に帰着される．

$$\text{maximize } \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t. } 0 \leq \alpha_i \leq C (i = 1, \ldots, l), \sum_{i=0}^{l} \alpha_i y_i = 0 \tag{2.10}$$

以上は線形分離可能な場合であるが，SVM はカーネルトリックにより非線形分離も可能である．非線形変換 Φ を用いてより高次元空間に写像しその高次元空間で線形分離を行うことで実質的な非線形分離を可能にする．ここで元の空間で定義され，Mercer の条件を満たすカーネル関数 $K(x, \dot{x})$ を導入することで，写像空間での複雑な計算を避けて元の空間で直接解くことができる．一般的なカーネル関数として，式 (2.11) 式 (2.12) で定義される Radial Basis Function(RBF) や Polynomial があり本研究ではこれを用いている．

$$K(x, \dot{x}) = \exp\left(-\frac{\|x - \dot{x}\|^2}{\sigma^2}\right) \tag{2.11}$$

$$K(x, \dot{x}) = (1 + x^T \dot{x})^p \tag{2.12}$$

こうして非線形分離可能な場合の目的関数は式 (2.13) のように書きかえられる．

$$\text{maximize } \sum_{i,j=1}^{l} -\frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.13}$$

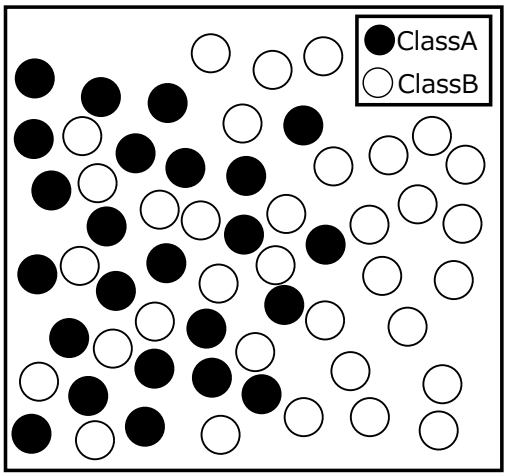
3 SVMにおける学習データ選択法

3.1 概要

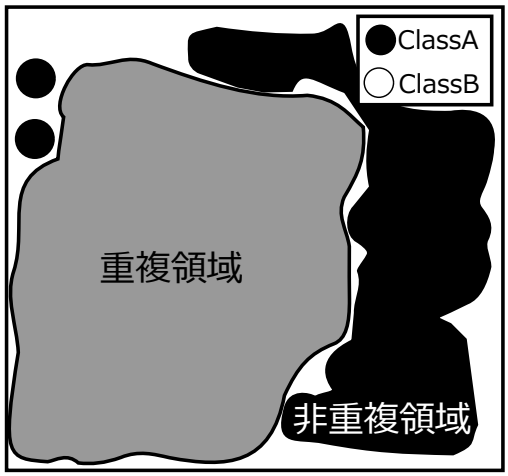
一般に実データは Fig. 2(a) のように完全な線形分離あるいは非線形分離が不可能な場合が多い．しかし通常の二値分類問題における SVM では classA と classB に分類する基準が学習されるため，誤識別が生じる．そこで，予測可能領域/予測不能領域（classA/予測不能領域あるいは classB/予測不能領域）に分類する基準を学習することで一方の class の学習事例に対して完全な分離を行う方法を提案する (Fig. 2(d))．

通常，分布が重複している領域を予測に使うことはできないが，重複していない領域であれば予測に使用できる．そこで重複している領域は無視して，重複していない領域を境界にする識別関数を引くことで特徴空間上の識別関数の片側領域（重複していない領域）を予測可能領域に分けられる．本手法では通常，与えられた全学習データを用いて行われる SVM の学習を，一方の class のデータを全て学習データとして用い，もう一方の class のデータの選択することで実現する．つまり Fig. 2(c) のように重複領域の片側の class のデータ (図では classB) を識別線の学習対象から除外し重複領域と非重複領域の境界への識別線を学習する．

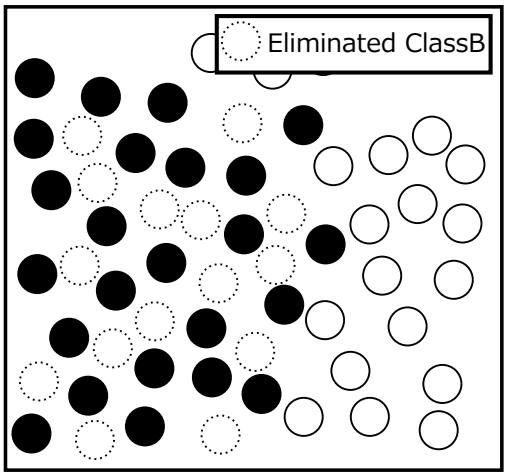
この場合，片側の class のデータを全て学習データとして用い，重複領域を無視することで，予測可能領域に対する完全な分離を保障している．しかしながら予測不可能に分布する classB 学習デー



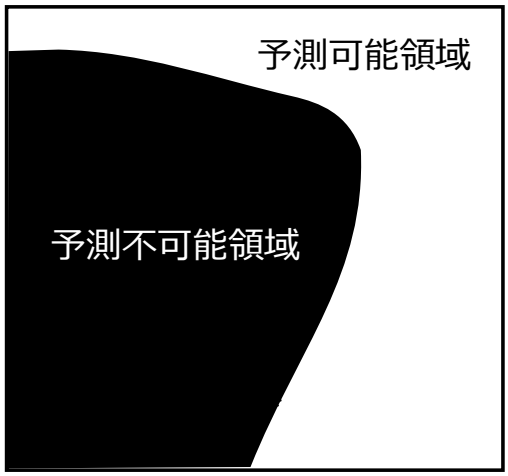
(a) 手順 1



(b) 手順 2



(c) 手順 3



(d) 手順 4

Fig. 3.1　提案手法の概要

タの中にはノイズとなるようなデータがあると考えられる．そこで予測不可能領域における誤識別の度合いを数値化し，その度合いに応じて予測可能領域を広くすることを検討する (Fig. 3)．

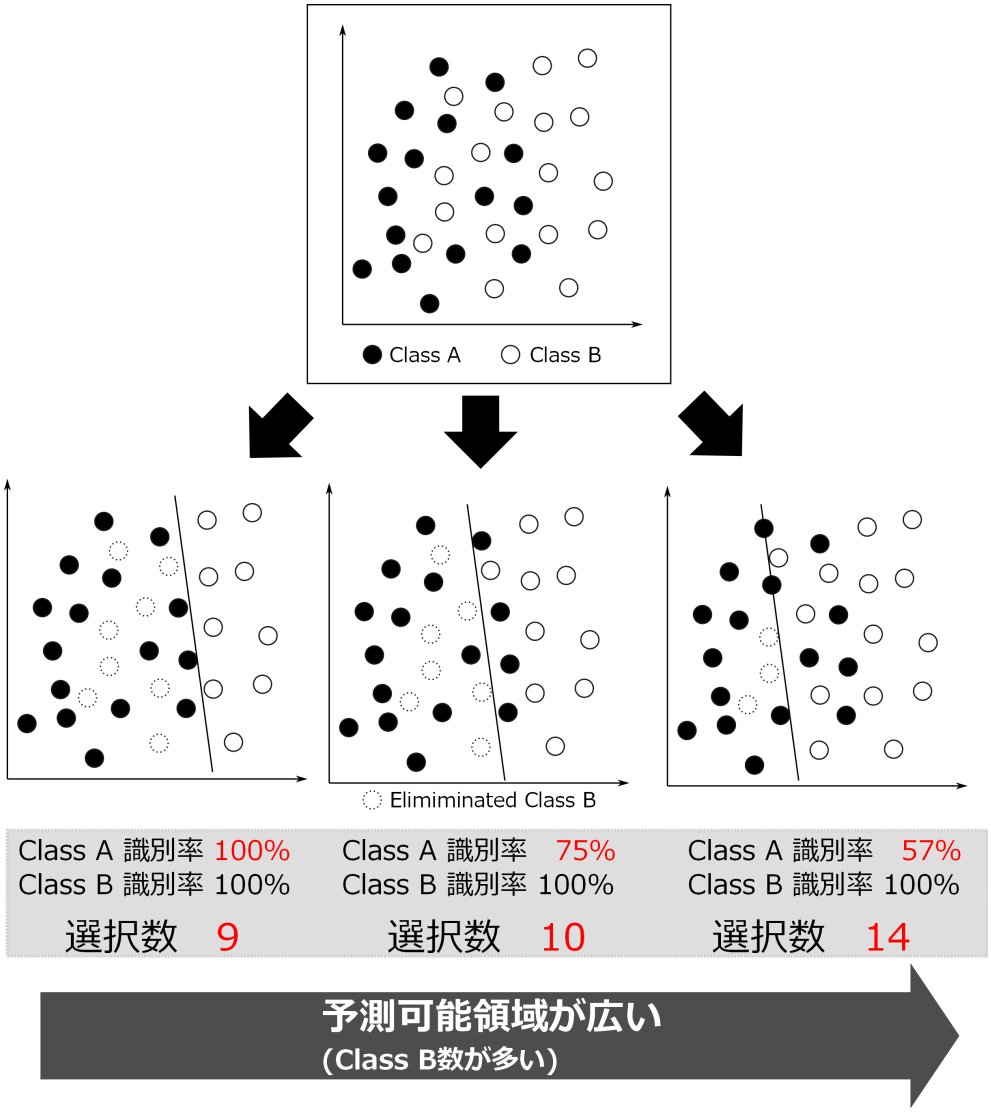


Fig. 3.2　誤識別の許容による予測可能領域の拡大

本手法の識別線の学習は，学習から除外される学習データの組合せを設計変数とし，次節で述べる目的関数を最大化する最適化問題を解くことで実現される．

3.2　提案手法の定式化

非重複領域と重複領域の境界に学習される識別関数の評価を行うために以下の 2 つの評価基準（目的関数）を考えた．学習データは classA と classB の 2 つのいずれかに属すると考える．classA のデータを全て用い，classA を選択して識別関数の学習を行い特徴空間を classA と判断される領域と判断不可能領域に分離する場合を考える．

O_1 識別線から誤識別される classB データまでの距離の総和 (SVM Confidence Margin)

O_2 正しく識別される classA の学習データ数

O_2 は予測可能領域を広くするため,O_1 はクラス B の誤識別を少なくするため（予測可能領域における誤識別をなくすため）の目的関数である．これら O_1,O_2 は競合関係にありそのトレードオフ関係に応じたパレートフロントを得るためにこの目的関数を採用した．

一つ目の目的関数 O_1 として識別面から誤識別される全 classB データまでの距離の総和を用いており，最小化を行う．実際には距離ではなく SVM Confidence Margin[8] を用いている．

学習データを特徴ベクトルと class のペア $(\boldsymbol{x}, y)$ で表し $y = \{-1, 1\}$ とする．学習データの集合 $\boldsymbol{X} = \{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_N}, y_N)\}$ によって学習される関数を $f_{\boldsymbol{X}}$ とするとき入力$\boldsymbol{x}$に対する出力は $f_{\boldsymbol{X}}(\boldsymbol{x})$ で表されるものとする．学習データ $(\boldsymbol{x}, y)$ に対する SVM Confidence Margin は $yf(\boldsymbol{x})$ によって表される．SVM Confidence Margin は正しく分類されるデータは正の値をとり，誤って分類されるデータは負の値をとる．

$$m(y, f(x)) = \begin{cases} 1 & \text{if } y = 1 \text{ and } yf(x) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

O_1 の最小化は式 (3.2) で表される．

$$\text{minimize } O_1 = \sum_{i=1}^{l} y_i f(x_i) m(y_i, y_i f(x_i)) \tag{3.2}$$

O_1 は誤って識別される classB（選択しない class）の学習データが多いほど値が大きくなる．またその誤識別から識別線が遠いほど値が大きくなる．つまり識別線から近い位置に誤りがあり，かつその誤り classB 数が少ないほど値が小さくなる．classB の誤識別がない場合，O_1 は最小値 0 になる．

2 つ目の目的関数 O_2 は，学習された識別関数によって正しく分類される classA のデータ数と誤って classA が分類される数の差を，基準として用いており，最大化を行う．学習される識別関数 $f(x)$ の正しい識別 $y = f(x)$ が

$$l(y, f(x)) = \begin{cases} 0 & \text{if } y \neq f(x) \\ 1 & \text{if } y = f(x) \end{cases} \tag{3.3}$$

で表される時，目的関数 O_2 の最大化は式 (3.4) で表される．

$$\text{maximize } O_2 = \sum_{q=1}^{k} l(y_q, f(x_q)) - \alpha \left(k - \sum_{q=1}^{k} l(y_q, f(x_q)) \right) \tag{3.4}$$

O_2 は正しく分類される classA(設計変数として選択している class) が多く，誤って分類される classA が少ないほど評価が良くなる．O_2 が良くなることは予測可能領域が広くなることを意味する．

3.3 NSGA-II による提案手法の実現

提案手法の学習データの選択を，多目的最適化問題としてとらえ，NSGA-II(Elitist Non-Dominated Sorting Genetic Algorithm)[9] により実現する方法を述べる．NSGA-II は 2001 年に Deb，Agrawal

らによって提案された．NSGA-II では，保存するための母集団（アーカイブ母集団）と交差・突然変異といった遺伝的操作を用いた探索を行うための探索母集団の 2 つの独立した母集団を用いた解探索を進めていく[10]．NSGA-II には，適合度の高い個体の保存，多様性に優れた個体の選択など，多目的 GA における重要な機構が組み込まれており，優れた探索性能を有することが報告されている．

3.3.1 解候補の遺伝子表現

学習データの組合せを表す N ビットストリングと SVM のパラメータである C と polynomial Kernel の P を設計変数とした．全学習データ数が N 個ある場合，解候補は N ビットストリング $(s_1 s_2 s_3 \ldots s_N)$ で表現される (Fig. 4)[11]．例えばデータ $\{x_i, y_i\}$ は $s_i = 1$ のとき学習データに含まれ，$s_i = 0$ のとき含まれない．また遺伝子型はそのまま表現型として使用する．

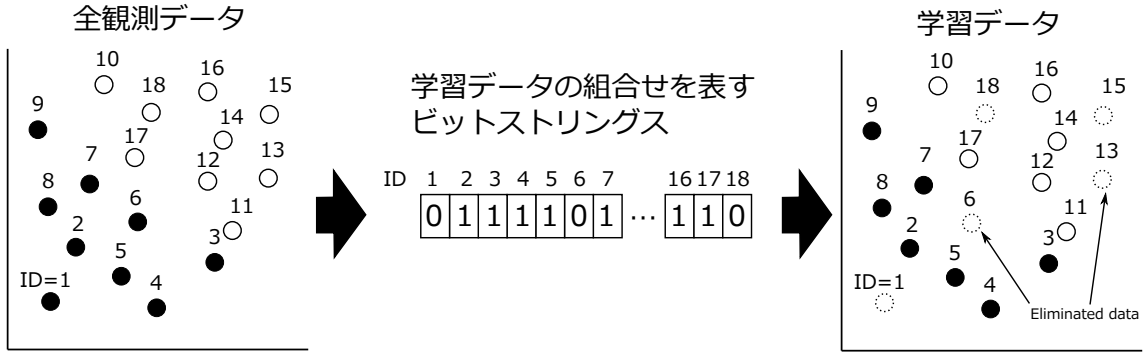


Fig. 3.3 解候補の遺伝子表現

3.4 評価実験

人工のデータセットを用いて提案手法のテストを行った．Table 5 に NSGA-II のパラメータを示す．

SVM のパラメータ C の範囲は 2^{-5} から 2^{15}[12]，次数 p は最大学習データ数の 10%以下の整数に設定した．学習は Soft Margin SVM の Polynomial Kernel および Linear Kernel によって行った．ただし Polynomial Kernel は p=2 に固定する場合と，p を 1 から 20 の範囲で変動させる 2 パターンのテストを行った．またテストでは classA の選択を行った．NSGA-II に適応するため式 (3.4) の目的関数 O_2 を式 (3.5) のように最小化問題に変換した．

$$\text{minimize} \quad O_2^{'} = \frac{1}{O_2} \tag{3.5}$$

人工データセットとして Fig. 5 のように分布する 2 次元のデータを用いた．データ数は 200 点で 2class（各 100 点) で構成される．

3.4.1 実験結果

Fig. 6 に Linear kernel を使用した時の NSGA-II の探索によって得られた全パレートフロントを示す．Fig. 7 に Polynomial kernel を p を 2 に固定して使用した時の結果を示す．Fig. 8 に Polynomial kernel を p の範囲を 1 から 20 とした時の結果を示す．

Linear kernel ではパレート解が 20 個，Polynomial Kernel p=2 では 18 個，Polynomial Kernel

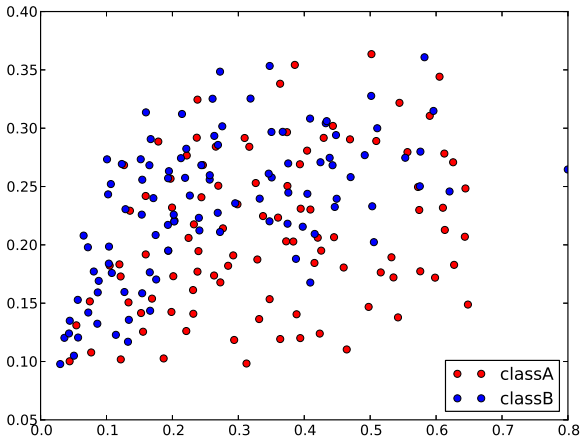


Fig. 3.4　人工データ：200 点

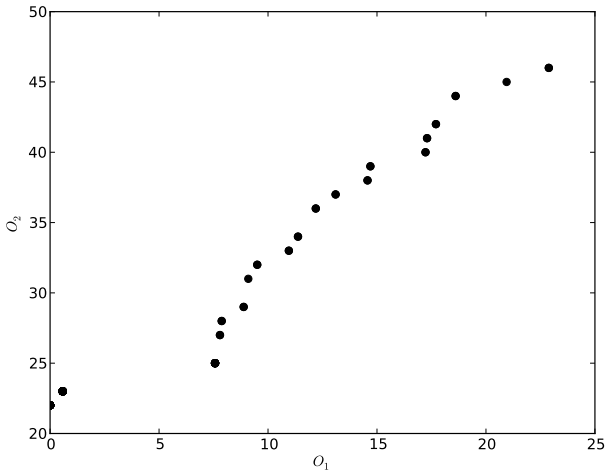


Fig. 3.5　パレートフロント（Linear kernel）

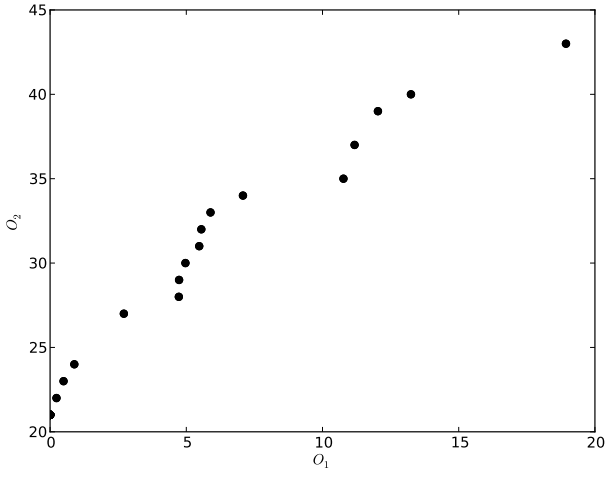


Fig. 3.6　パレートフロント（Polynomial Kernel p=2）

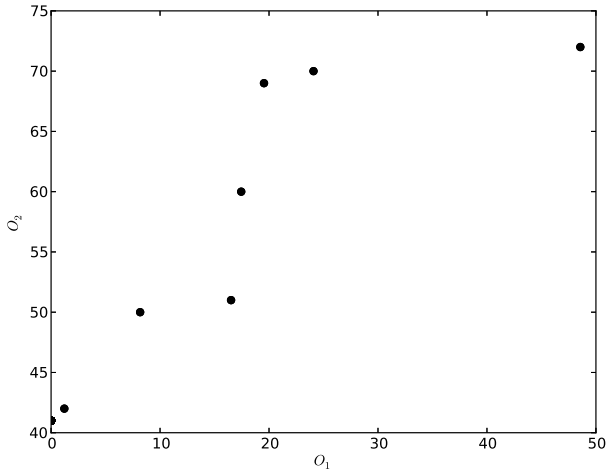


Fig. 3.7　パレートフロント（Polynomial Kernel p=1-20）

p=1-19 では 10 個のパレート個体が得られた.Fig. 9 に得られたパレート解について，その選択された学習データと，学習された識別関数の目的関数の特徴空間上の分布を示す．Fig. 9 のパレート解 1 が評価値 O_2 が高く，評価値 O_1 の評価が低いパレート解の分布であり，パレート解 4 が評価値 O_2 が低く, 評価値 O_1 の評価が高いパレート解の分布である．Fig. 10 に Polynomial Kernel の p を 2 に固定して解析した結果を示す．Fig. 11 に Polynomial Kernel の p を 1 から 20 の間で解析した結果を示す．Table 5 に得られたパレート解について選択された学習データの最大値，最小値，平均値を記載する．Table 5 に得られたパレート解について FN 率 (False Negative Rate) の最大値，最小値，平均値を記載する．

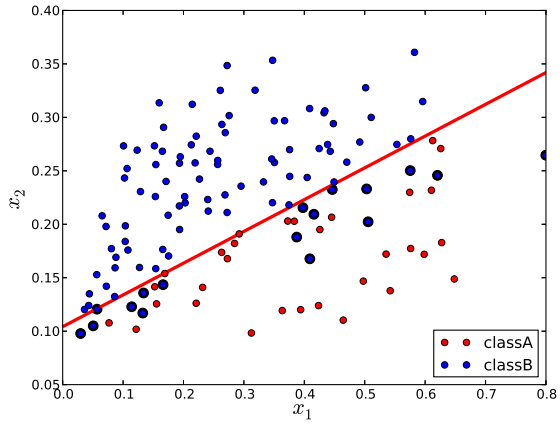
Table 5 において全てのカーネルにおいて最小 FP 率 (False Positive Rate) であるが 0 が求まっていることから識別線を境界にして予測可能領域と予測不可能領域に分離できていることが確認できた．また Fig. 9，Fig. 10，Fig. 11 のパレート解 4 の分布をみることからも識別線を境界にして上側が予測不可能領域，下側が予測不可能領域に分かれていることが確認できた．

4　ユーザインタフェースの開発

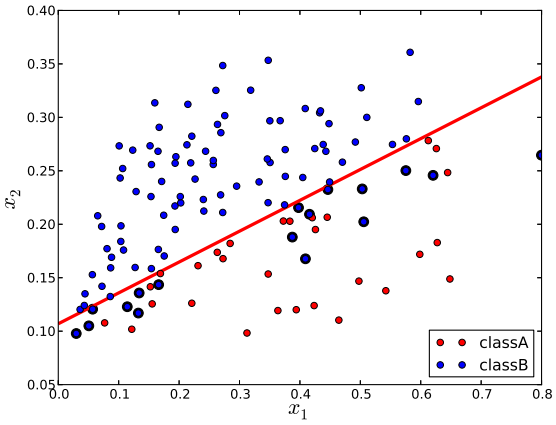
4.1　システムの概要

3 章で提案した SVM の利用法によって学習データに対する誤差が異なる識別基準が学習されるようになった．ここでは得られた複数の識別基準の中から妥当なものを検討するためのインタフェースを開発する．

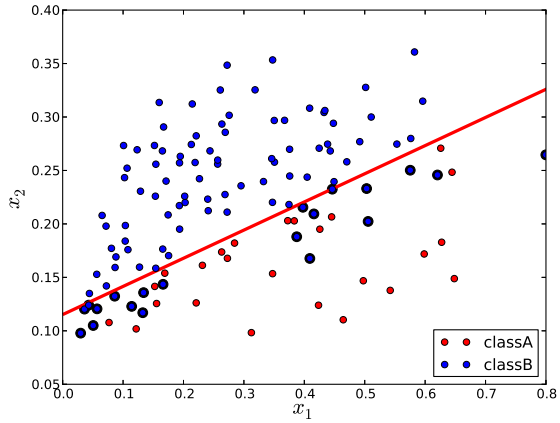
本システムは提案手法によって得られる複数の識別基準を視覚的に検討するためにパラメータの違いによる識別線と学習データの分布を一覧表示する機能を搭載する．また異なるカーネルやパラメータによる NSGA-II の解析結果を比較するために解析結果をデータベースにおいて一元的に管理する機能を搭載する．学習データに対する識別率やサポートベクトル数などの SVM 解析結果に応じて表示画像を並べ替える機能を搭載する．さらに指定した誤差内に収まる結果を一目でわかるように表示



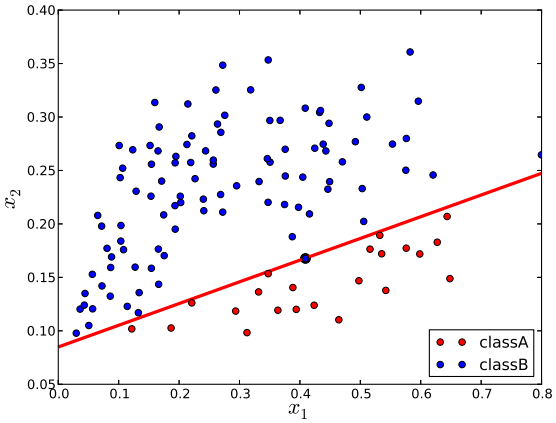
(a) パレート解 1



(b) パレート解 2

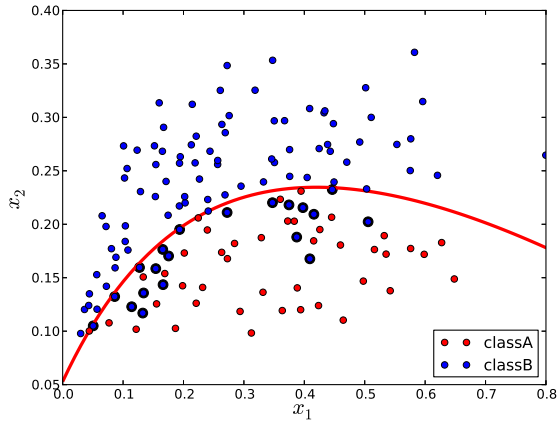


(c) パレート解 3

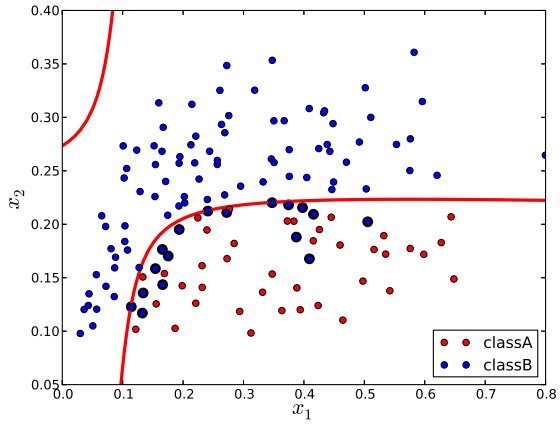


(d) パレート解 4

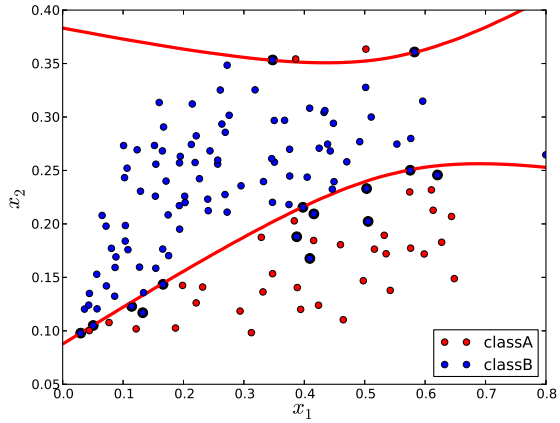
Fig. 3.8　選択された学習データと学習された識別関数 (Linear Kernel)



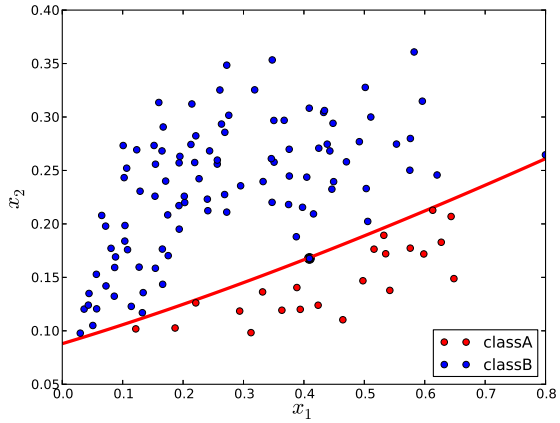
(a) パレート解 1



(b) パレート解 2

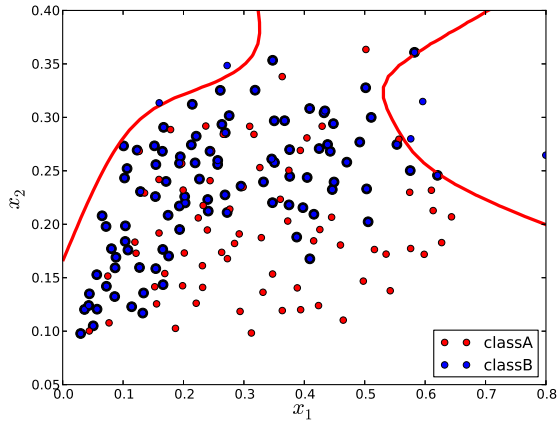


(c) パレート解 3

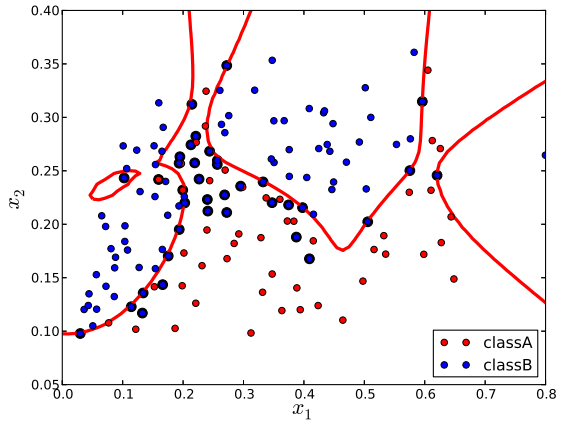


(d) パレート解 4

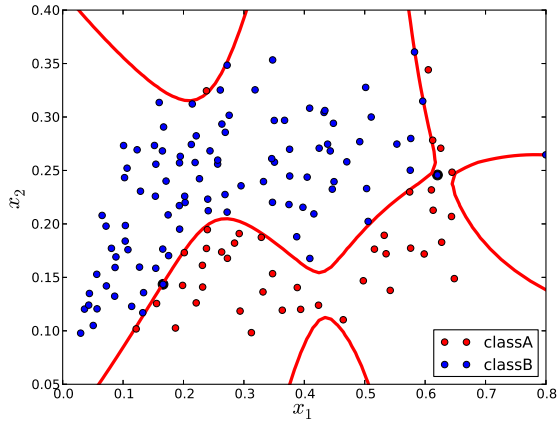
Fig. 3.9　選択された学習データと学習された識別関数 (Polynomial Kernel P=2)



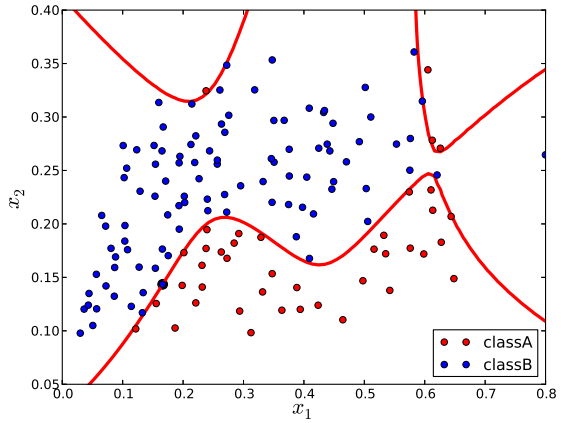
(a) パレート解 1



(b) パレート解 2



(c) パレート解 3



(d) パレート解 4

Fig. 3.10　選択された学習データと学習された識別関数 (Polynomial Kernel P=1-20)

する機能を搭載することで，ユーザが求める誤差に収まる抽出したり，試行錯誤的にユーザが求める誤差をしたりできるようにする．これらの機能により解析結果によって得られた識別基準の検討が可能になることが期待される．

4.2　ユーザインタフェースとデータベース

本システムは大きく，パレート解情報の入力から，その情報を基にした SVM による解析，2 次元の分布図の作成，解析結果のデータベースへの格納までを自動的に実行する機能と，データベースに登録された識別基準を表示するビュワー機能から構成される．本システムのユーザインタフェースを Fig. 12 に示す．

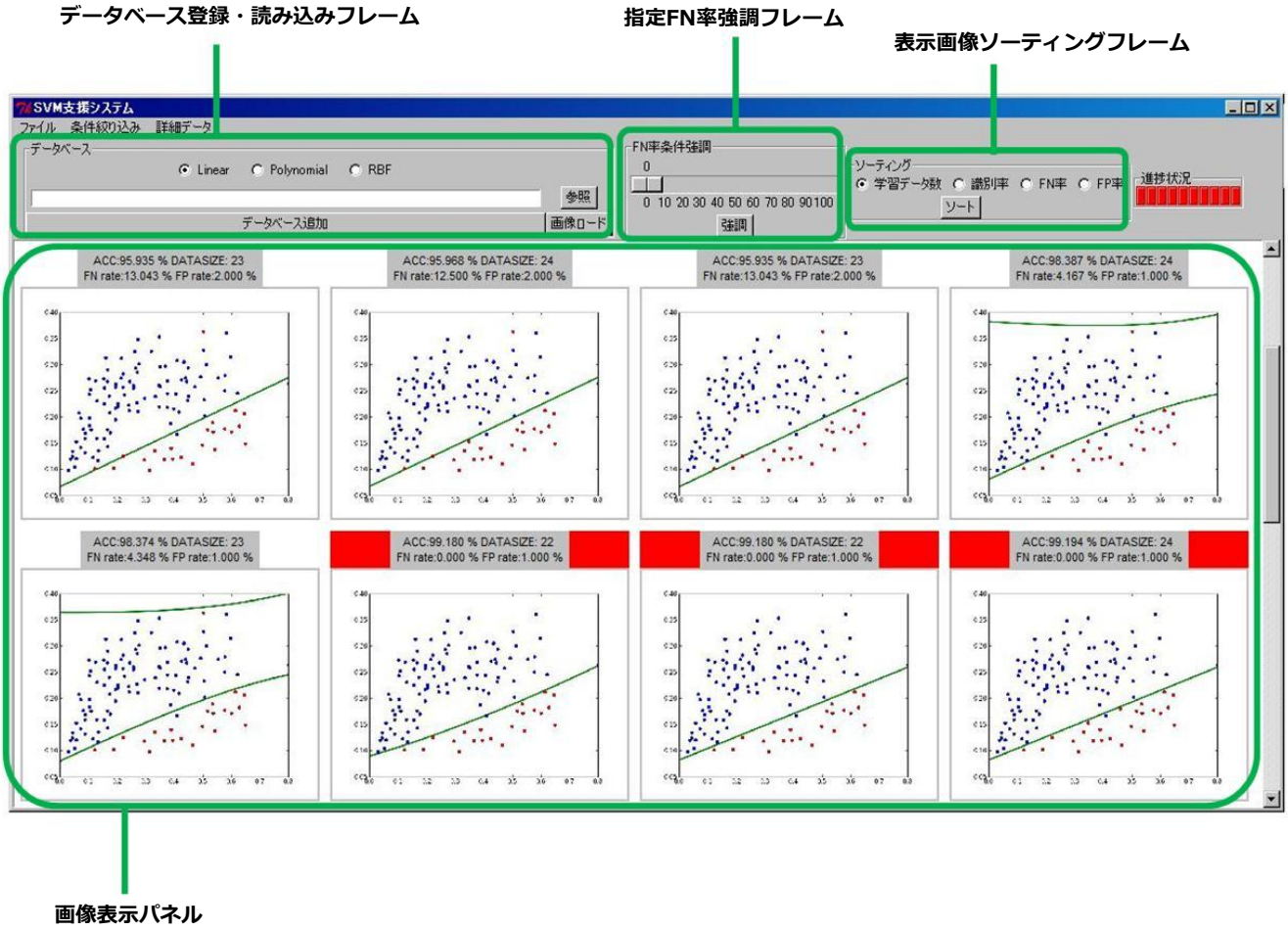


Fig. 4.1　インタフェース

インタフェースはデータベースの登録，読み込みを行うデータベース登録・読み込みフレーム，読み込んだ情報を表示する画像表示パネル，画像表示パネルに表示される画像の並べ替えを行う表示画像ソーティングフレーム，特定の結果を強調する指定 FN 率強調フレームから構成される．

解析結果はリレーショナルデータベースに格納する．ここでは，格納用のデータベースとして SQLite を用いた．データベースのテーブルは Table 5 に示す構造を持つ．主キーはデータセットの種類，カーネルの種類及び，NSGA-II の設計変数情報である．

4.3 ユーザインタフェースの利用方法

NSGA-II 解析結果のデータベースへの格納

NSGA-II の解析によって学習データに対する誤差を許容するに応じて，予測可能領域が拡大するトレードオフ関係にある結果が得られる．解析結果を一元的に管理するために解析結果をデータベースに登録する．

具体的な操作としてはデータベース登録・読み込みフレームのラジオボタンで登録する学習結果の SVM のカーネルの種類を選択し，設計変数情報を記述したファイルのパスを「参照」ボタンもしくはテキストボックスで指定し「データベース追加」ボタンを押すことで設計変数情報をもとに学習が行われ，学習結果がデータベースに登録される．

登録されている解析結果の一覧表示

データベース登録・読み込みフレームの画像ロードボタンをクリックすることでデータベースに登録されている全学習結果の学習データ分布と識別線を描写した画像が画像表示パネルに一覧表示される．また画像と同時にデータベースに登録されている属性の一覧表示も行う．

表示画像の並べ替え

表示画像ソーティングフレームのラジオボタンで属性を指定しソートボタンをクリックすることで画像表示パネルに表示された画像が指定した属性で並べ替えられる．

指定する FN 率以下の画像の強調表示

指定した誤差内に収まる結果を赤枠で表示することで，解析データに適当な誤差を求たり，任意の誤差に収まる結果を抽出する．

具体的な操作としては指定 FN 率強調フレームのスクロールバーで FN 率を指定し,「強調」ボタンをクリックすることで画像表示パネルに指定 FN 率以下の識別線の背景が赤色で強調して表示される．

5 結論

薬効が既知の患者の情報が与えられたとき SVM によって未知の患者に対する薬効を予測するシステムの実現にあたり，分離不可能問題を対象として予測可能領域を抽出する SVM の利用技術を考案した．これは特徴空間を予測可能領域と予測不能領域に分類する基準を学習することで一方の class の学習事例に対して誤識別のない識別基準を求め限定的な予測を試みる方法である．この学習は重複領域の片側の class を識別線の学習対象から除外し，重複領域と非重複領域の境界へ識別線を学習することで実現する．また予測不能領域に分布する class の中にはノイズとなるようなデータがあると考え予測不能領域における誤識別の度合いを数値化しその度合いに応じて予測可能領域を広くすることを考慮した．

本手法は，学習から除外される学習データの組合せを設計変数とし，誤識別されるデータまでの距離の総和と正しく識別される学習データ数を目的関数とする最適化問題を解くことで実現される．NSGA-II に提案手法を適用し，分離不可能問題に対して実験を行った結果，予測可能領域と不可能領域に分離する基準が学習され，かつ予測可能領域が拡大するにつれて予測可能領域における誤識別率が向上するトレードオフ関係に応じた識別基準集合が得られることが確認された．

また NSGA-II の解析結果によって得られる結果をデータベースに格納し，識別基準を視覚的に検討するためのインタフェースの開発を行った．

謝辞

本研究を遂行するにあたり，多大なる御指導そして御協力を頂きました，同志社大学生命医科学部の廣安知之教授に心より感謝いたします.

本研究を進める上で，多くの助言と丁寧なご指導を頂きました，同志社大学生命医科学部の横内久猛教授に心より感謝いたします.

また本論文を執筆するあたり，校正してくださました山中亮典さんに感謝いたします．授業や論文執筆，就職活動などで忙しい中，丁寧な校正や助言をしていただきありがとうございました.

データマイニンググループの一員としてミーティングにおいて多くの助言や指摘をしていただきました横田山都さん，西井琢真さん，大堀裕一さんに感謝いたします.

またドクターの田中美里さん，ソーシャルウェアグループの宮地正大さんからも私の研究に関して，鋭い指摘やアドバイスをしていただきありがとうございました.

最後に私が研究室において活動する上で，精神的，経済的にサポートし続けてくだった両親に感謝して，修士論文といたします.

参考文献

1) 大内香. 抗体医薬の現状と展望
-トラスツズマブを例に-. 日本薬理学雑誌, Vol. 136, No. 4, pp. 210–214, 2010.

2) Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.

3) BISHOP C. M. パターン認識と機械学習上. ベイズ理論による統計的予測, 2007.

4) TSUDA Koji. Overview of support vector machine. *The Journal of the Institute of Electronics, Information, and Communication Engineers*, Vol. 83, No. 6, pp. 460–466, 2000.

5) 津田宏治. サポートベクターマシンとは何か. 電子情報通信学会誌, Vol. 83, No. 6, pp. 460–466, 2000.

6) Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In Claire Nedellec and Celine Rouveirol, editors, *Machine Learning: ECML-98*, Vol. 1398 of *Lecture Notes in Computer Science*, pp. 137–142. Springer Berlin / Heidelberg, 1998.

7) E. Osuna, R. Freund, and F. Girosit. Training support vector machines: an application to face detection. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pp. 130 –136, 1997.

8) Ling Li, Amrit Pratap, Hsuan-Tien Lin, and Yaser Abu-Mostafa. Improving generalization by data categorization. In *Knowledge Discovery in Databases: PKDD 2005*, Vol. 3721 of *Lecture Notes in Computer Science*, pp. 157–168. Springer Berlin / Heidelberg, 2005.

9) K. Deb, S. Agarwal, A. Pratap, and T. Meyarivan. A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II. In *KanGAL report 200001, Indian Institute of Technology, Kanpur, India*, 2000.

10) 渡邉真也. 遺伝的アルゴリズムによる多目的最適化に関する研究, 2003.

11) 柳浦睦憲, 茨木俊秀. 組合せ最適化問題に対するメタ戦略について (情報基礎理論ワークショップ (la シンポジウム) 論文小特集). 電子情報通信学会論文誌. D-I, 情報・システム, I-情報処理, Vol. 83, No. 1, pp. 3–25, 2000.

12) Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering,National Taiwan University,Taipei, 2003.

付 図

付 表

Table 1　Parameter of NSGA-II and SVM

最大世代数	1000
突然変異率	1/遺伝子長
交叉率	1.0
アーカイブサイズ	100
母集団数	100

Table 2　選択された学習データ数

	Linear	Polynomial(P=2)	Polynomial(P=1-20)
best	46.0	43.0	72.0
worst	22.0	21.0	41.0
average	34.6	30.5	52.5

Table 3　False Positive Rate

	Linear	Polynomial(P=2)	Polynomial(P=1-20)
best	0.00	0.00	0.00
worst	0.24	0.2	0.95
average	0.17	0.08	0.38

Table 4　リレーショナルデータベースのテーブル構造

Primary key	Attribute	Type	Meanings
○	kerneltype	文字列型	SVM カーネルの種類．”L”,”P”,”G”によってそれぞれ線形カーネル，多項式カーネル，RBF カーネルを表す
○	traincombin	文字列型	学習データの組合せをデータセットのデータサイズ長のビットストリングで表す．1 : 学習する 0 : 学習しない
○	C	浮動小数点型	ソフトマージン SVM のコストパラメータ C
○	P	浮動小数点型	RBF カーネルのパラメータ σ
○	D	整数型	多項式カーネルの次数 p
○	dataname	文字列型	データセット名
	path	文字列型	識別線と学習データの画像を保存している場所の path
	acc	浮動小数点型	学習データに対する識別率
	fp_r	浮動小数点型	学習データに対する False Positive Rate
	fn_r	浮動小数点型	学習データに対する False Negative Rate
	trainN	整数型	学習データ数
	ms	浮動小数点型	マージンサイズ