修士論文

パターン認識と最適化を用いた病理データの分類法の検討

同志社大学大学院 生命医科学究科 生命医科学専攻

博士前期課程 2012 年度 1024 番

大堀　裕一

指導教授　廣安知之教授

2014 年 1 月 24 日

Abstract

In this paper, a methodology that combines a classification algorithm with a combinatorial optimization problem is proposed.Using the proposed algorithm, classification problems of medical data such as breast cancer data can be performed more accurately.Using the conventional method, the breast cancer data is classified into malignant growth and benign tumor.However, by the conventional method, the incorrect discernment exists in many cases.To conquer the defect, the proposed method classify the breast cancer data into two groups; (1) the data which definitely belongs to malignant growth area, and (2) the data which has the possibility to belong to either malignant growth area or benign tumor area. This method can be attained by solving the combination optimization problem of learning data.Moreover, in this method, judgment is possible only in the limited area.Then, it is considered extending the area which can be judged by permitting incorrect discernment.This problem can be attained by solving a multi-objective optimization problem about the combination of learning data.The classification algorithm based on learning data selection is proposed, and the effectiveness of the proposed algorithm is discussed through the numerical experiments.

目 次

1　序論

癌の診断において，患者の腫瘍が悪性であるか良性であるかという判断は病理診断によって行われる．病理診断は，病理学の知識や医師の経験によって診断を行う専門性の高い技術であり，明確な基準がなく診断する医師によって結果が異なるといった問題点が存在する．そこで過去の病理診断によって得られた患者の腫瘍のデータを使用して，医師がある患者の腫瘍が良性か悪性か診断する際に参考となる情報を提示する診断支援ツールが求められている．しかし，悪性腫瘍と良性腫瘍のデータの分布に重複する領域がある場合，良性と悪性の領域に分けても両方の領域において誤分類が存在し，患者の腫瘍データが入力された際良性，悪性どちらであるか正確に判断できない．そこで良性，悪性という２つのクラスに完全に分離することが困難なデータに対して，良性または悪性であると確実に判断できる領域とどちらか判断できない領域に分ける事を提案する．すると，今まで正確に判断できなかったデータに対しても，一部領域においては確実に判断することが可能となる．しかし，この方法では一部の領域でしか判断できず，結果が判断できない場合が存在する．この問題を解決するため医師に判断可能領域においてある程度の誤識別を許容した場合の識別結果を提示できるよう，誤識別の度合いに応じて判断可能領域を拡大することの検討を行う．本研究では病理データとして患者の乳房から採取した乳癌データを用いる．

本稿では，2 章で病理診断の現状と問題点，3 章でパターン認識を用いた病理データの分類，4 章で分類法の提案，5 章で提案した分類法の評価実験，6 章で結論を述べる．

2　病理診断の現状と問題点

病理診断とは，生体から摘出した組織を薄く切りだして染色し，医師が顕微鏡で観察することによって病変の有無や種類を診断する技術である．癌の診断においても，患者の腫瘍が良性であるか悪性であるかという判断は最終的には病理診断によって行われている．Fig.1 は病理画像の例である．病理診断の現状として，医師は病理学の知識や自らの経験によって診断を行い主観的に判断している．診断を行う際医師は，病理画像から特徴量を抽出し，その数値をもとに良性であるか悪性であるかという指標を用いることをせず，顕微鏡による目視のみによって判別している．そのため明確な診断基準がなく，診断する医師によって結果が異なるといった問題点が存在する．こうしたことから，過去の病理診断によって得られたデータをもとに，診断の際の指標となる情報を提供するシステムが求められている．

3　病理データによる病態の分類方法

3.1　病理画像の特徴量を用いた分類

本研究では，人体から摘出した腫瘍の病理画像から得られたデータを使用している．このデータは，ある腫瘍が良性であるか，悪性であるかというクラスと，その腫瘍の病理画像から抽出したいくつかの特徴量を値として持つ．ある病理画像から得られたデータは特徴空間と呼ばれる d 次元空間の 1 点として表現できる．Fig.2 は 2 次元特徴空間の例である．この d 次元ベクトルを特徴ベクトルと呼ぶ．上記のデータを良性と悪性の 2 クラスに分類する問題として，入力と出力間の関数を，与えられたデータから学習する方法を考える．学習とは学習データを利用して，未知の特徴ベクトルがどちらのクラスに属するか判定する関数を求めることである．本研究では，患者の腫瘍病理画像の特徴量を入力とし，その腫瘍が良性であるか悪性であるかということを出力とする関数を学習することを考える．学習は特徴空間上で 2 クラスの識別線を決定することに相当し，パターン認識という考え方を用いる．

3.2　パターン認識

パターン認識とは，認識対象がいくつかの概念に分類できるとき，観測されたパターンをそれらの概念のうちの 1 つに対応させる考え方である．パターン認識におけるこの概念をクラスと呼ぶ．パターン認識では，未知のデータを正しく分類することが目標となる．

n 個の観測データ $\{\vec{x}_i, y_i\}, i = 1, \ldots, n$ が与えられているとする．このとき，$\vec{x}_i \in \vec{R}^n$ 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クラスである．また，関数 $f : \vec{R}^n \to R$ が次の条件を満たすものとする．

$$f(\vec{x}_i) > 0 \quad \text{if } y_i = 1$$
$$f(\vec{x}_i) < 0 \quad \text{if } y_i = -1$$

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 $\vec{x}$ に対応するクラス y を

$$y = \mathrm{sgn}(f(\vec{x})) \tag{3.1}$$

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

$$\mathrm{sgn}(f(\vec{x})) = \begin{cases} 1 & \text{if } f(\vec{x}) \geq 0 \\ -1 & \text{if } f(\vec{x}) < 0 \end{cases} \tag{3.2}$$

によって表される符号関数である．

3.3 Support Vector Machine

3.3.1 線形 SVM

SVM[1) 2)] ではクラスを分類する超平面を求めることによって，未知データの分類を行う．超平面 H_0 が式 (3.3) によって表されるとする．

$$H_0 : \vec{w} \cdot \vec{x} + b = 0 \tag{3.3}$$

ここで，$\vec{w}$ は超平面の法線ベクトルであり，b は定数項である．d_+ と d_- を超平面から最も近い正と負のサンプルまでの最短距離とするとき，超平面のマージンは $d_+ + d_-$ となる．与えられたデータが線形分離可能な場合，SVM はマージンが最大となる超平面を求める．これは以下のように定式化される．まず，次の制約条件を満たすものとする．

$$\vec{x}_i \cdot \vec{w} + b \geq +1 \quad \text{for } y_i = +1 \tag{3.4}$$

$$\vec{x}_i \cdot \vec{w} + b \leq -1 \quad \text{for } y_i = -1 \tag{3.5}$$

これらの条件は式 (3.6) にまとめることができる．

$$y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \forall i \tag{3.6}$$

式 (3.4) が成り立つ点は，超平面 $H_1 : \vec{x}_i \cdot \vec{w} + b = 1$ 上に存在する．同様に，式 (3.5) が成り立つ点は，超平面 $H_2 : \vec{x}_i \cdot \vec{w} + b = -1$ 上に存在する．このとき，$d_+ = d_- = 1/||\vec{w}||$ であるから，マージンは $2/||\vec{w}||$ となる．したがって，式 (3.6) の下で $||\vec{w}||^2$ を最小化することによってマージンを最大化する問題として，SVM は式 (3.7) に示すように定式化できる．

$$\begin{aligned} &\text{minimize} \quad ||\vec{w}||^2 \\ &\text{subject to} \quad y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \forall i \end{aligned} \tag{3.7}$$

2 次元の特徴空間における超平面の例を Fig.3 に示す．Fig.3 において，H_1 と H_2 上に位置する丸で囲まれた点（式 (3.6) の等号が成り立つ点）をサポートベクターと呼び，これらが取り除かれた場合には異なる超平面が得られる．

次に，Lagrange 関数を用いることによって，より扱いやすい双対問題へと帰着させる．まず，式 (3.6) の各制約条件に対して Lagrange 乗数 $\alpha_i,\ i = 1, \ldots, l,\ (\alpha_i \geq 0)$ を定義する．これより，式 (3.7) の Lagrange 関数を

$$L(\vec{w}, b, \vec{\alpha}) \equiv \frac{1}{2}||\vec{w}||^2 - \sum_{i=1}^{l} \alpha_i \{y_i(\vec{x}_i \cdot \vec{w} + b) - 1\} \tag{3.8}$$

とする．このとき，式 (3.7) を式 (3.8) を用いて書き換えると次のようになる．

$$\text{minimize} \quad \max_{\alpha \geq 0}\{L(\vec{w}, b, \vec{\alpha})\} \tag{3.9}$$

この問題の双対問題は次のようになる．

$$\begin{aligned} \text{maximize} \quad & \min_{x}\{L(\vec{w}, b, \vec{\alpha})\} \\ \text{subject to} \quad & \alpha_i \geq 0 \quad i = 1, \ldots, l \end{aligned} \tag{3.10}$$

なお，式 (3.10) の最小化問題の最適解では L の勾配が 0 になるため

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.11}$$

$$\vec{w} = \sum_{i=1}^{l} \alpha_i y_i \vec{x}_i \tag{3.12}$$

となる．したがって，式 (3.10) の問題は次のように書き換えることができる．

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j \\ \text{subject to} \quad & \alpha_i \geq 0 \quad i = 1, \ldots, l \\ & \sum_{i=1}^{l} \alpha_i y_i = 0 \end{aligned} \tag{3.13}$$

3.3.2 非線形 SVM

パターン認識では，線形 SVM では分離することが不可能な場合も存在する．そのような場合，SVM では式 (3.14) に示すような写像 Φ を用いることによって，サンプルデータをより高次元の空間 $\mathcal{H}$ に写し，$\mathcal{H}$ 上で線形分離することが提案されている [3) 4) 5)]．

$$\Phi : \vec{R}^n \mapsto \mathcal{H} \tag{3.14}$$

写像 Φ を使うことにより，識別関数は

$$f(\Phi(\vec{x})) = \vec{w} \cdot \Phi(\vec{x}) + b \tag{3.15}$$

$$= \sum_{i=1}^{l} \alpha_i y_i \Phi(\vec{x}_i) \cdot \Phi(\vec{x}) + b \tag{3.16}$$

となる．このとき，

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j) \tag{3.17}$$

となるカーネル関数 K が存在する場合には，Φ の計算を行う必要がない．したがって，カーネル関数 K を用いることによって識別関数は式 (3.18) のように表される．

$$f(\vec{x}) = \sum_{i=1}^{l} \alpha_i y_i K(\vec{x}_i, \vec{x}) + b \tag{3.18}$$

主なカーネル関数の種類として，線形 SVM の場合は線形カーネル，非線形 SVM の場合は多項式カーネル，ガウスカーネルが存在する．線形カーネルは式 (3.19)，多項式カーネルは式 (3.20)，ガウスカーネルは式 (3.21) のように表される．

$$K(x_i, x_j) = x_i^T x_j \tag{3.19}$$

$$K(x_i, x_j) = (r + \gamma x_i^T x_j)^d, \gamma > 0 \tag{3.20}$$

$$K(x_i, x_j) = \exp\left(-g||x_i - x_j||^2\right), g > 0 \tag{3.21}$$

3.4 SVM による病理データの分類

過去の病理診断によって得られた病理画像から特徴量を抽出し，特徴空間で表すと Fig.2 のような分布になるとする．このような特徴空間に対して，上記で説明した SVM により良性と悪性に分類すると Fig.4 のような識別線で良性領域と悪性領域に分けられる．すると，新たな患者の病理データが入力された際，良性領域に入力されたとすると良性と判断でき，悪性領域に入力されたとすると悪性と判断できる．このように SVM では過去の病理診断によって得られた病理データの特徴空間の分布を学習することで識別線を算出し，この識別線をもとに新患者の病態を分類できる．

4 最適化を用いた病理データ分類法の提案

4.1 病理データの３クラス分類

4.1.1 3 クラス分類の概要

Fig.2 のような，良性データと悪性データが重複する部分が存在する学習データの場合，完全なクラスの分離が困難であり，従来の方法で良性と悪性のクラスに分けたとしても誤った分類が存在してしまい，新たな患者の病理データが入力された際，診断支援システムの信頼性が低下する．そこで，悪性と良性のデータが重複した領域は，クラスの判断が不可能な領域として，クラスが重複していない判断可能領域 (悪性または良性のデータのみが存在する領域) とクラスが重複している判断不可能領域の 3 クラスに分ける手法を提案する．これによって，一部領域において良性または悪性であるか過去のデータからは正確に判断が可能となる．さらに，判断可能領域を広くすることが求められる．これは，良性または悪性のみが存在する領域をできる限り大きくするように学習データの組み合わせを選択することで実現する．例えば，悪性と判断できる領域と判断不可能な領域に分離したい場合，悪性の学習データの組み合わせを選択する．Fig.5 は悪性のデータを判断可能領域として選択した例である．するとクラスの分離が不可能であったデータが，Fig.5 のように過去のデータから悪性と判断できる領域と，どちらか判断できない領域に分離することができる．同様に，良性と判断できる領域と，どちらか判断できない領域に分離することも可能である．学習データ選択の方法としてはデータの組み合わせ最適化問題と捉え，最適化を行った．最適化アルゴリズムには，組み合わせ最適化問

題に適した遺伝的アルゴリズム (Genetic Algorithm:GA)[6) 7) 8)] を用いた．GA については以下で説明する．

4.1.2　遺伝的アルゴリズム

ここでは最適化アルゴリズムである GA について説明する．GA は生物が環境に適応して進化していく過程を工学的に模倣した確率的な最適化手法である．自然界における生物の進化過程においては，ある世代を形成している個体集団の中で環境に適応した個体がより高い確率で生き残る．ここで生き残った個体が次世代に子を残す．この生物進化のメカニズムをモデル化し，環境に対して最もよく適応した個体，すなわち目的関数に対して最適値を与えるような解を計算機上で求めることが GA の概念である．GA では 1 つの解候補を 1 個体として扱い，個体の集団を用いて解探索を行う．解候補は設計変数からなるベクトル表現や構造体など，問題によって異なる表現をとる．個体は設計変数値をコーディングした染色体により特徴づけられる．そして，染色体をベクトル表現や構造体にデコーディングし，目的関数値を計算する．なお，染色体は複数の遺伝子で構成されている．各個体は目的関数値に応じた適合度を有し，ある世代を形成している個体集団において，適合度の高い個体ほど高確率で生き残るように選択を行う．加えて，交叉および突然変異といった個体生成の遺伝的操作によって子個体を生成し，次世代の個体集団を形成する．この世代更新の繰返しによって適合度の高い個体が集団内に増加し，最適解に集団を収束させるのが GA のメカニズムである．GA の基本的な流れを以下に示す.

初期化（Initialization）初期母集団を形成する複数の個体をランダムに生成し，各個体の適合度を評価する.

終了判定（Terminate Check）あらかじめ定められた終了条件に基づいて，GA の処理を終了する．この時，母集団で適合度の最も高い個体を最適解として採用する．一般的には，世代数による終了条件が使用される.

評価（Evaluation）各個体に環境に合わせた適合度，すなわち目的関数値を計算する.

複製選択（Selection of Parerents）次世代の子を生成するための親個体候補を選択する.

交叉（Crossover）親個体 A の遺伝子と親個体 B の遺伝子を入れ替えることにより新しい子個体を生成する.

突然変異（Mutation）染色体上のある遺伝子を突然変異率に従って他の対立遺伝子に置き換える.

生存選択（Selection of Survivals）交叉，突然変異によって生成された子個体から，次世代に残る個体を選択する.

4.1.3　3 クラス分類の定式化

n 個のデータ $(\boldsymbol{x_i}, y_i)$,$i = 1...n$ 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クラスである．設計変数を学習データの組み合わせとして，選択するデータを 1，選択しないデータを 0 でデータ長の 2 値ビット配列で表し，

Fig.6 のように学習データに対応させる．n 個の中から k 個が選択されたとき，学習データ数は k 個となる．また，与えられた学習データのクラスと，SVM による分類が異なれば誤識別となる．誤識別は式 (4.1) のように定義される．

$$l(y, f(\boldsymbol{x})) = \begin{cases} 1 & \text{if } y \neq s(f(\boldsymbol{x})) \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

学習データ数が k 個の場合，学習データに対する誤識別率 Err は式 (4.2) のように表される．

$$Err = \frac{1}{k} \sum_{i=1}^{k} l(y_i, f(\boldsymbol{x}_i)) \tag{4.2}$$

$f(x)$ は識別関数の値，$s(f(x))$ は識別関数の符号を表す．

判断可能としたいクラスのデータを式 (4.3) で表す．判断可能領域の広さを式 (4.3) で表されるデータの数で表現すると，このデータ数が大きくなるほど領域は広くなる．そこで，式 (4.5) で表される選択された学習データの誤識別率を 0 にするという制約条件で，式 (4.4) で表される目的関数 O を最大化する．制約条件を満たさない場合はペナルティとして減じ，誤識別率が 0 となるようにする．すると最終的に良性または悪性のみが存在する領域を分類することができる．

$$m(y, f(\boldsymbol{x})) = \begin{cases} 1 & \text{if } y = 1 \text{ and } s(f(\boldsymbol{x})) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

$$\text{maximize } O = \sum_{i=1}^{k} m(y_i, f(\boldsymbol{x}_i)) \tag{4.4}$$

$$\text{subject to } Err = 0 \tag{4.5}$$

4.2 判断可能領域の拡張

4.2.1 判断領域拡張法の概要

前節では病理データの 3 クラス分類の方法を提案し，良性または悪性のみが存在する一部の領域において正確な判断ができる手法を提案した．しかし，新たな患者の病理データが上記以外の判断不可能領域に入力された際，診断支援ツールでは判断不可能と判定されることになる．そこで，ここでは判断不可能領域として捉えるのではなく判断領域における誤識別データを許容することで，判断できる領域を拡張した病理データの分類を提案する．Fig.7 のように誤識別データの度合いに応じた識別線を算出することで，良性または悪性である確率の等高線を誤識別度合いに応じて求めることができる．こうすることで，3 クラス分類では判断不可能領域としていた領域に新たな患者の病理データが入力された際も，診断支援ツールは良性または悪性の可能性を数値として出力することができる．
ここで，Fig.8 のように判断領域の誤識別の度合いが大きくなるにつれ判断領域が拡大され，誤識別度合いを小さくすると判断領域は縮小することが分かる．このことから判断領域の大きさと誤識別の

度合いはトレードオフ関係にあることが分かる．そこでこの問題を多目的最適化問題として考える．多目的最適化問題とはトレードオフ関係にある複数の目的をそれぞれ最適化する問題である．本手法では多目的最適化アルゴリズムとして MOEA/D(Multiobjective Evolutionary Algorithm Based on Decomposition)[9] を使用した．

4.2.2 判断領域拡張の定式化

n 個のデータ $(\boldsymbol{x_i}, y_i)$,$i = 1...n$ が与えられているとする．3 クラス分類の場合と同様に，設計変数を学習するデータの組み合わせとしてデータ長の 2 ビット配列として表す．n 個の中から k 個が選択されたとき，学習データ数は k 個となる．ここでは 2 つの目的関数を考える．まず 1 つ目に判断できる領域を最大化することである．判断できる領域の最大化としては，判断したい領域において正しく識別されたデータ数を最大化することで実現する．判断したい領域のデータは式 (4.6) で表すことができる．すると 1 つ目の目的関数は式 (4.7) で表すことができる．

$$c(y, f(\boldsymbol{x})) = \begin{cases} 1 & \text{if } y = 1 \text{ and } f(\boldsymbol{x}) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

$$\text{maximize } O_1 = \sum_{i=1}^{k} c(y_i, f(\boldsymbol{x}_i)) \tag{4.7}$$

2 つ目に誤識別の度合いを最小化することである．誤識別の度合いとしては，判断したい領域における誤識別データの識別線からの距離の総和として表す．判断したい領域における誤識別データは式 (4.8) で表される．すると誤識別データの識別線からの距離の総和は式 (4.9) で表すことができる．

$$d(y, f(\boldsymbol{x})) = \begin{cases} 1 & \text{if } y = 1 \text{ and } s(f(\boldsymbol{x})) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

$$\text{minimize } O_2 = \sum_{i=1}^{k} y_i f(x_i) d(y_i, y_i f(x_i)) \tag{4.9}$$

上記の 2 つの目的関数のパレート解を求めることによって，判断したい領域における誤識別の度合いに応じた分類を行うことができる．

5 評価実験

5.1 実験方法

本実験では，病理データとして UCI Machine Learning Repository の breast cancer[10] を用いた．このデータは乳房組織から腫瘍を採取したものであり，569 個のデータが存在する．データの形式としては腫瘍が良性であるか悪性であるかというラベルを示し，腫瘍の病理画像から抽出した 10 個の特

徴量を情報としている.[11)]．今回の実験で用いる特徴量としてはSVMによる識別率が最も高くなったtextureとconcave pointsの2つを使用した．Fig.9に2次元データの分布を示す．またbreast cancerデータの特徴量をTable1に示す.

5.1.1　3クラス分類の実験

ここでは，従来の2クラス分類の識別結果と3クラス分類による識別結果を比較する．実験に使用したSVMとGAのパラメータをそれぞれTable2,Table3に示す．以下に実験の手順について説明する.

手順 1 textureとconcave pointsによる2次元データに対してスケーリングを行い正規化する.

手順 2 2次元データに対してSVMにより良性領域と悪性領域の2クラスに分類する.

手順 3 2次元データに対して提案した3クラス分類を行い，良性領域，悪性領域，判断不可能領域を求める.

手順 4 2クラス分類の結果に対して3-fold cross validationを行い，未知データに対する識別結果を評価する.

手順 5 3クラス分類に対して3-fold cross validationを行い，未知データに対する識別結果を評価する.

5.1.2　判断領域拡張の実験

ここでは，判断領域拡張について評価実験を行う．実験に使用したMOEA/Dのパラメータ Table4に示す．以下に実験の手順について説明する.

手順 1 textureとconcave pointsによる2次元データに対してスケーリングを行い正規化する.

手順 2 提案した判断領域拡張を行い，誤識別の度合いに応じた良性領域，悪性領域の分類を行う.

手順 3 手順2で分類した領域に対してそれぞれSVMのパラメータをcが10^{-3}-10^{10}，dが1-10の範囲において最適なパラメータを探索する．ただし，パレート解の関係を崩さないように制約を与える.

手順 4 誤識別の度合いに応じた領域分類の結果ごとに3-fold cross validationを行い，未知データに対する識別結果を評価する.

5.2　実験結果

5.2.1　3クラス分類の結果

2クラス分類による結果をFig.10に，3クラス分類による結果をFig.11に示す．3-fold cross validationによりそれぞれの未知データに対する誤識別数を評価するとFig.12のような結果となった．このことから3クラス分類を行うことで，新たな患者の病理データが入力された際，誤った分類を減らすことができると考えられる.

5.2.2　判断領域拡張の結果

判断領域拡張の結果を Fig.13-Fig.18 に示す．また，Fig.19 のようなパレート解集合が得られた．3-fold cross validation によりそれぞれの未知データに対する誤識別数と識別率を評価すると Fig.20, Fig.21 のような結果となった．許容する誤識別データの数が大きくなるほど未知データに対する誤識別数は多くなっているが，識別率が高くなっているため，判断不可能と出力される可能性が低くなると考えられる．このことから判断領域の拡張を行い，許容する誤識別数に応じた分類を行うことで，3 クラス分類では判断不可能とされる場合においても，良性，または悪性である確率を出力することができる．

6　結論

本稿では，医師が病理診断を行う際の補助となる情報を提示するため，患者の腫瘍の病理画像から得られたデータを学習して，新たに診断を行う患者の腫瘍が良性であるか，悪性であるか識別する方法を提案した．しかし，良性と悪性の 2 つのクラスに完全に分離することが困難な場合，SVM で 2 つのクラスに分けたとしても誤識別が存在し，診断の信頼性が低下する．そこでこういったデータに対して，一方のクラスのデータを選択し，良性または悪性であるか過去のデータから判断が可能な領域と，判断不可能な領域の 3 クラスに分ける手法を提案した．この問題はデータの組み合わせ最適化問題と捉え，学習データの組み合わせを設計変数とし，誤識別が 0 という制約条件で学習データ数を最大化する最適化問題とすることで実現した．従来の 2 クラス分類と未知データに対する誤識別数を比較すると，3 クラス分類において誤識別数が減少し，誤った診断を減らす可能性が示唆された．しかし，この方法では判断不可能と出力される場合が存在するため，誤識別を許容することで判断領域を拡張させることについても提案した．この方法は，許容する誤識別データ数と判断領域において正しく識別されるデータ数がトレードオフ関係にあることを考え，多目的最適化を行った．その結果，誤識別数に応じた領域を分類することができた．この結果に対し評価実験を行ったところ，3 クラス分類では判断不可能であった場合でも識別結果を返すことができ，その有効性が示唆された．

謝辞

謝辞

参考文献

1) Vladimir Vapnik, "The support vector method of function estimation", in J.A.K. Suykens and J.Vandewalle Nonlinear Modeling, Advanced Black-Box Techniques, Kluwer Academic Publishers, Boston, pp.55-85, 1998.

2) Vladimir Vapnik, "Statistical learning theory", John Wiley, New York, 1998.

3) Terrence Furey, Nello Cristianini, Nigel Duffy, David Bednarski, Michel Schummer and David Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data", Bioinformatics, vol.16, no.10, pp.906-914, 2000.

4) Sujun Hua and Zhirong Sun, "Support vector machine approach for protein subcellular localization prediction", Bioinformatics, vol.17, no.8, pp.721-728, 2001.

5) Nello Cristianini, John Taylor, "An Introduction to Support Vector Machines:And Other Kernel-Based Learning Methods", Cambridge University Press, 2000.

6) David Goldberg, "Genetic algorithms in search, optimization and machine learning", Addison Wesley, 1989.

7) Darrel Whitley, "A genetic algorithm tutorial", Statistics and computing, vol.4, no.2, pp.65-85, 1994.

8) Ting Chen, Chung Chen, "Improvement of simple genetic algorithm in structural design", International journal for numericalL method in engineering, vol.40, pp.1323-1334, 1997.

9) Qingfu Zhang, Senior Member, IEEE, and Hui Li, "MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition", IEEE Transactions on Evolutionary Computation, vol.11, no.6, pp.712-731, 2007.

10) UCI Machine Learning Repository. [Online]. Available: `http://archive.ics.uci.edu/ml/datasets.html`

11) Nick Street, William Wolberg, Olvi Mangasarian, "Nuclear Feature Extraction For Breast Tumor Diagnosis", International Symposium on Electronic Imaging, Science and Technology, vol.1905, pp.861-870, 1993.

付 図

付 表

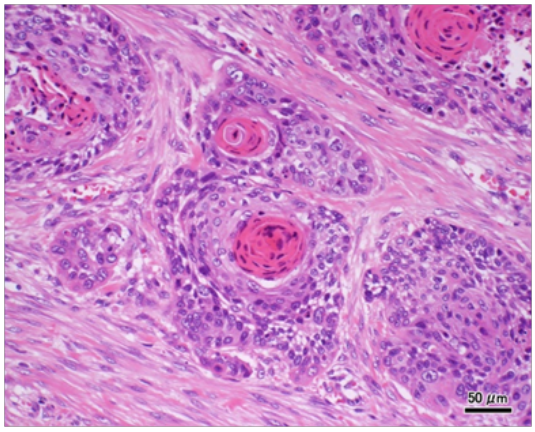

Fig. 1　食道癌の病理画像 (出典:日本病理学会)

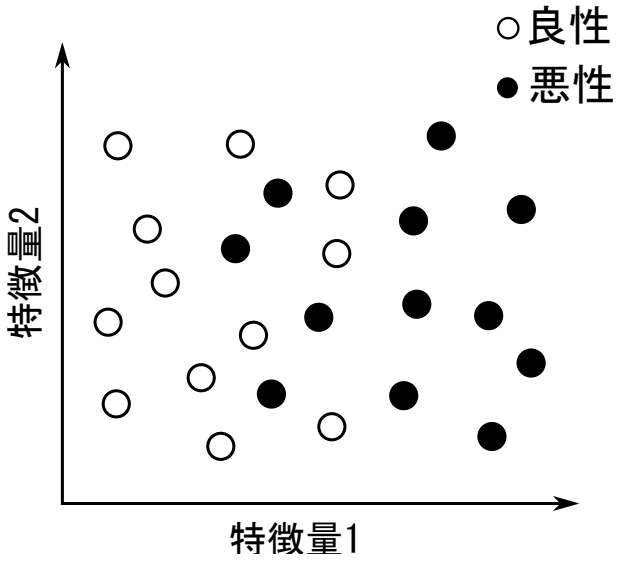

Fig. 2　2 次元特徴空間の例

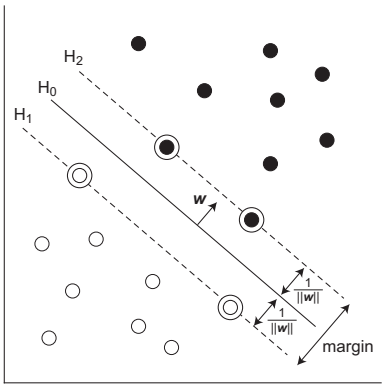

Fig. 3　2 次元特徴空間における超平面

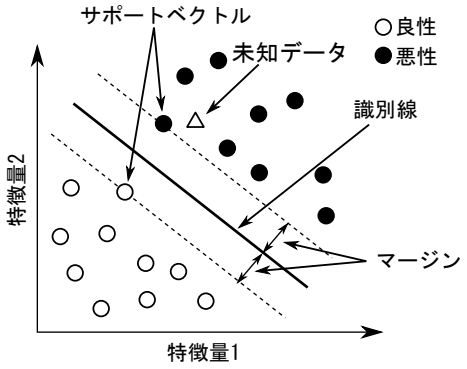

Fig. 4　SVM による良性悪性の分類

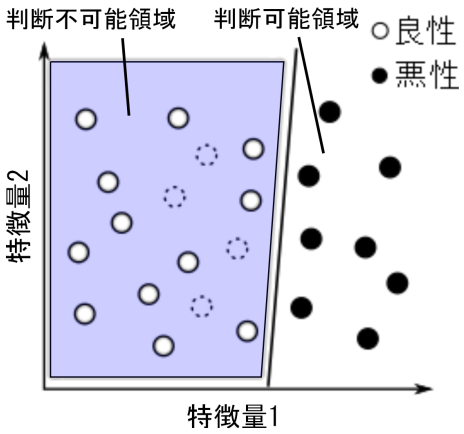

Fig. 5　悪性データ選択の例

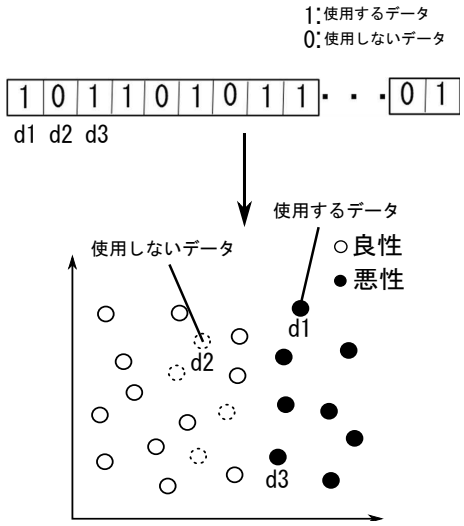

Fig. 6　設計変数の表現方法

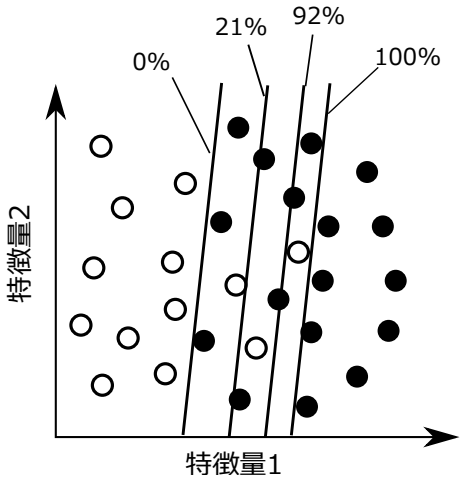

Fig. 7　誤識別度合いに応じた識別線

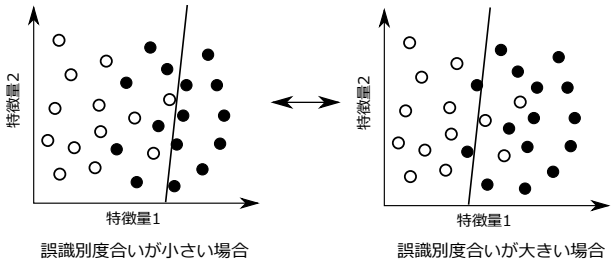

Fig. 8　誤識別度合いと判断領域の関係

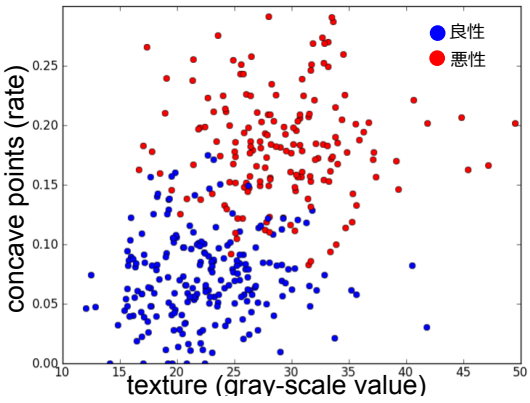

Fig. 9　乳癌データの分布図

Table 1　乳癌データの特徴量

	特徴量
1	radius
2	texture
3	perimeter
4	area
5	smoothness
6	compactness
7	concavity
8	concave points
9	symmetry
10	fractal dimension

Table 2　SVM のパラメータ

パラメータ	値
SVM の種類	C-SVM
カーネル関数	多項式
d	2
c	10000

Table 3　GA のパラメータ

パラメータ	値
世代数	300
個体数	100
選択手法	トーナメント選択
トーナメントサイズ	4
交叉方法	一様交叉
交叉率	0.9
突然変異率	0.01

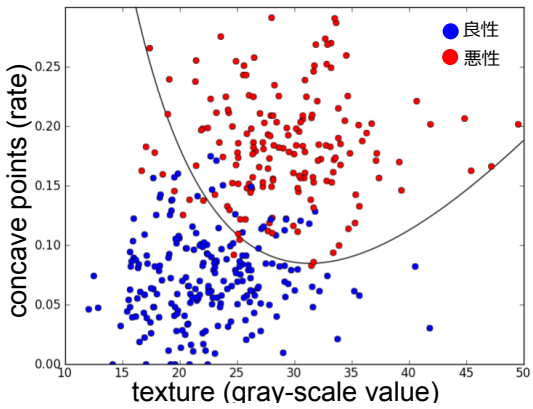

Fig. 10　2 クラス分類の結果

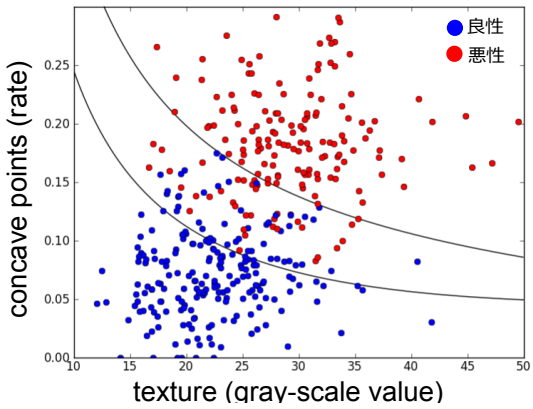

Fig. 11　3 クラス分類の結果

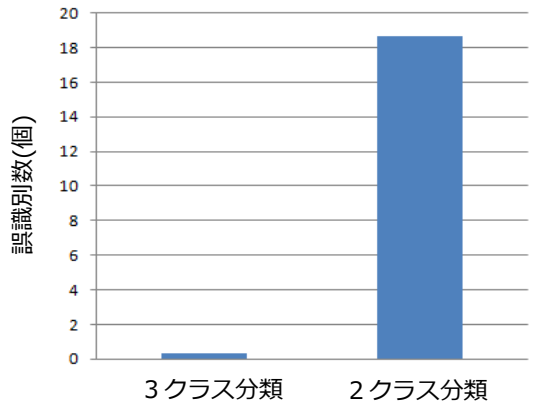

Fig. 12　2 クラス分類と 3 クラス分類の比較

Table 4　MOEA/D のパラメータ

パラメータ	値
世代数	200
個体数	300
交叉率	1.0
交叉方法	一様交叉
突然変異率	0.01
近傍距離	3

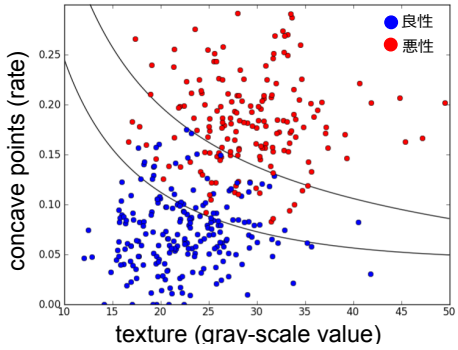

Fig. 13　誤識別を許容しない場合

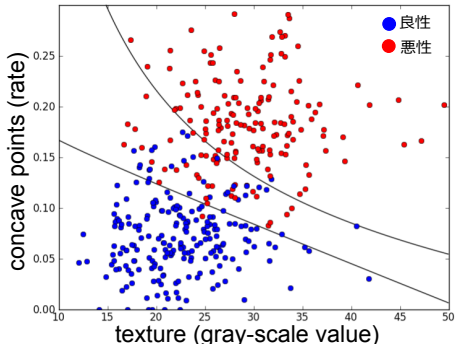

Fig. 14　誤識別を 1 個許容した場合

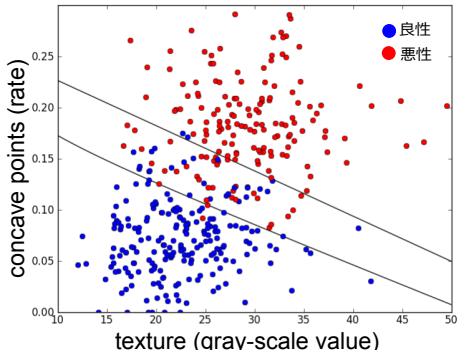

Fig. 15　誤識別を 2 個許容した場合

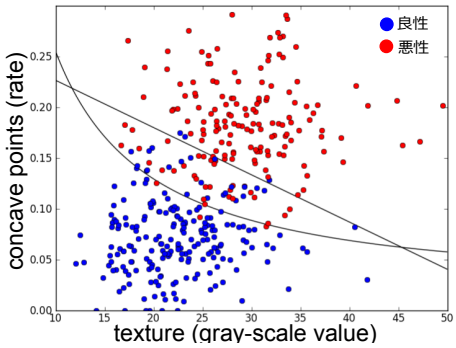

Fig. 16　誤識別を 3 個許容した場合

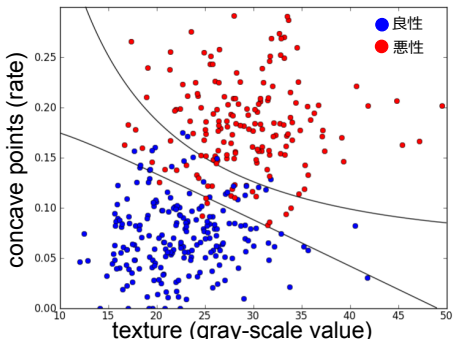

Fig. 17　誤識別を 4 個許容した場合

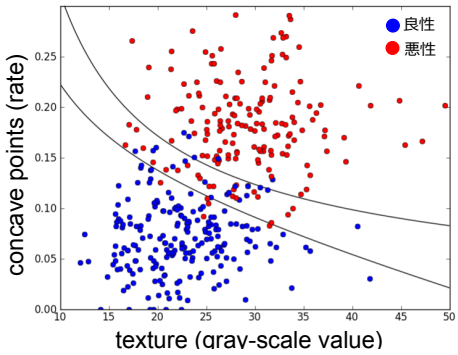

Fig. 18　誤識別を 5 個許容した場合

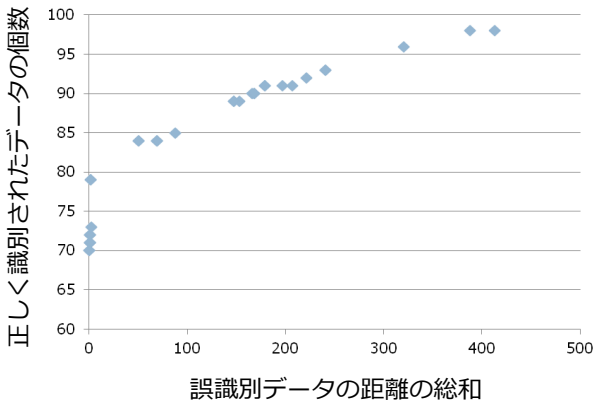

Fig. 19　判断領域拡張のパレート解

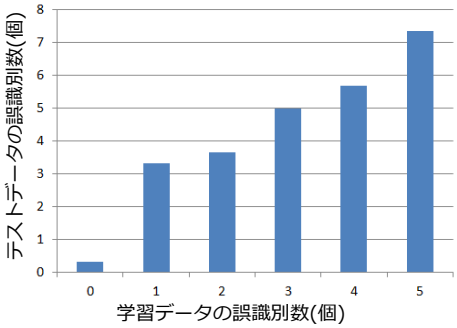

Fig. 20　分類結果ごとの誤識別数

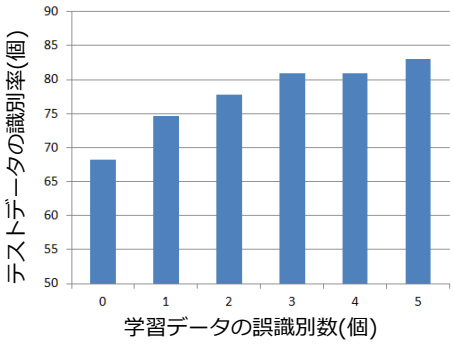

Fig. 21　分類結果ごとの識別率