修士論文

２クラス分類の為の多目的遺伝的プログラミングを用いた特徴量変換手法の検討

同志社大学大学院 生命医科学研究科
医工学・医情報学専攻　医情報学コース
博士前期課程 2013 年度 1034 番

白石　駿英

指導教授　廣安 知之教授

2015 年 1 月 23 日

Abstract

In this paper, a method for feature transforming the input space into one-dimensional space in order to improve the classification accuracy of the classifier. Making an optimal feature transformation, it is possible to obtain a high spatial characteristics of identification accuracy than the given feature space. In order to obtain this feature conversion function, the Genetic Programing (GP) is used. A conversion formula is represented by tree structure of GP. The newly proposed evaluation function in GP in contrast with previous studies have succeeded in finding a high-precision conversion function. On the other hand, if there is a bias in the number of data between classes, classification accuracy of a class which has a lot of data is high, but that of the other class is low in two-class classification. To solve this problem, based on the existing research, it was investigated how to handle a multi-objective optimization for this problem. Moreover, in this paper we investigated examination of multi-objective optimization techniques and the method how to determine the identification threshold line. In the results of the study, it became possible to obtain the transformation function which has higher classification accuracy of the two classes than existing methods.

目 次

1　序論

機械学習による識別[1)] は非常に有効であり，近年様々な応用が為されてきた．食品の金属成分の識別や癌の良性・悪性の識別といった問題等に適用されている[2)3)4)]．特にそのニーズは様々な場面に存在し，一例として癌腫瘍の診断がある．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システムが求められている．システムは病理画像によって得られた特徴量をもとに，パターン認識によって腫瘍の良性と悪性を識別する．

そのような識別を行う際には，パターン認識器がより良い識別を行うことを考慮しなければならない．学習器は代表的にニューラルネットワーク，Support Vector Machine や決定木等の様々な学習器が研究されている[5)6)7)]．しかし，それらは学習過程についての考察であり，データが保有している特徴量に対する前処理を考えたものではない．そこで，学習器の識別精度を向上させる為に，入力する特徴量空間に対する前処理が学習器による学習と独立して考えられる．前処理として高次元特徴量空間をより識別に有効な低次元特徴量空間に変換することが挙げられる．本稿では特徴量空間を変換する変換式を最適化する為，遺伝的プログラミング[8)] を用いる．遺伝的プログラミングによって識別に有効な特徴量へ変換し，良い識別精度を得たとする既存研究も報告されている．[9)10)11)]．本稿では特に，クラス間でデータ数に偏りのあるデータに対してバランスの良い識別を考慮する為，多目的遺伝的プログラミングを用いた特徴量変換式の最適化に対する検討を行う．

本稿では，2 章でパターン認識器と識別，3 章で遺伝的プログラミングを用いた特徴量変換の最適化，4 章で多目的遺伝的プログラミングによる特徴量変換の検討，5 章で結論を述べる．

2　パターン認識器と識別

2.1　パターン認識の概念

パターン認識[1)] 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 2 クラスの場合の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

$$y = \mathrm{sgn}(f(\vec{x})) \tag{2.1}$$

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

$$\mathrm{sgn}(f(\vec{x})) = \begin{cases} 1 & \text{if } \ f(\vec{x}) \geq 0 \\ -1 & \text{if } \ f(\vec{x}) < 0 \end{cases} \tag{2.2}$$

によって表される符号関数である．

2.2　代表的な識別器と識別

2.2.1　ニューラルネットワーク

代表的な識別器には，ニューラルネットワーク（Neural Network：NN），決定木（Decision Tree：DT）や SVM（Support Vector Machine：SVM）などがある[5) 6) 7)]．NN は人間の神経回路を模した学習器である．識別を行う為には，入力信号における神経細胞の反応を定義する活性化関数 $g(h)$ やその結合加重 w を定義し，再帰的な学習を行うことによって，最適なネットワークが構築される．ニューラルネットワークの基本構造であるパーセプトロンについて述べる．活性化関数 $g(h)$ の例（シグモイド関数）や活性化関数が処理する入力 h は式 (2.3)，式 (2.4) で定義される．

$$g(h) = \sigma(\beta h) = \frac{1}{1 + e^{-\beta h}} \tag{2.3}$$

$$h = \sum_{k=0}^{K} w_k x_k \tag{2.4}$$

またネットワークの学習について単層のパーセプトロンの場合，Fig.1 のように，入力層と出力層から成る．入力ユニット k から出力ユニット i への結合の強さを表す結線 w_{ik} が付され，入力層にはバイアス w_{i0} に対応する入力ユニット $x_0 = 1$ を用意する．入力層の入力に対して出力層が重みによる演算処理と活性化関数による演算処理が行い，出力が返される．その出力に対して，入力に対する出力の誤差を調べるため，式 (2.5) の 2 乗和誤差を算出する．

$$E = \frac{1}{2}\sum_{\mu=1}^{N}(f^{\mu} - y^{\mu})^2,\ f^{\mu} = w^T x^{\mu} \tag{2.5}$$

パーセプトロンはこの 2 乗和誤差を最小にする為に，最適重みを更新する．重みの修正を行う手法として再急降下法がある．最小化するべき目的関数を 2 乗和誤差関数 E として重みは式 (2.6) で更新される．

$$w_{t+1} = w_t + \Delta w_t \quad , \quad \Delta w_t = -\eta_t \frac{\partial E}{\partial w}|_{w_t} \tag{2.6}$$

この更新を学習回数分繰り返すことにより最終的なネットワークが構築される．この機構がニューラルネットワークの学習の基本となる．

2.2.2 Support Vector Machine

SVM(Support Vector Machine)[12) 13)] は入力特徴量空間をクラス分類する最適な超平面を求めることによって識別を行う．SVM は特に 2 クラス分類において優れた学習器である．超平面 H_0 は式 (2.7) によって表される．

$$H_0 : \vec{w} \cdot \vec{x} + b = 0 \tag{2.7}$$

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識別関数の出力が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．制約条件は以下の式 (2.8) で表される．

$$y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \forall i \tag{2.8}$$

制約条件の境界面で超平面のマージンは $d_+ = d_- = 1/||\vec{w}||$ となり，合計してマージンは $2/||\vec{w}||$ となる．従って，式 (2.8) 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，ハードマージン SVM は式 (2.9) に示されるように定式化される．

$$\begin{aligned} &\text{minimize} \quad ||\vec{w}||^2 \\ &\text{subject to} \quad y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \forall i \end{aligned} \tag{2.9}$$

2 次元の特徴量空間における超平面の例を Fig.2 に示す．Fig.2 において，H_1 と H_2 上に位置する丸で囲まれた点（式 (2.8) の等号が成り立つ点）をサポートベクターと呼ぶ．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 (2.8) 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 (2.10) の Lagrange 関数を

$$L(\vec{w}, b, \vec{\alpha}) \equiv \frac{1}{2}||\vec{w}||^2 - \sum_{i=1}^{l} \alpha_i \{y_i(\vec{x}_i \cdot \vec{w} + b) - 1\} \tag{2.10}$$

とする．このとき，式 (2.9) を式 (2.10) を用いて書き換えると次のようになる．

$$\text{minimize} \quad \max_{\alpha \geq 0} \{L(\vec{w}, b, \vec{\alpha})\} \tag{2.11}$$

この問題の双対問題は次のようになる．

$$\begin{aligned} \text{maximize} \quad & \min_{x} \{L(\vec{w}, b, \vec{\alpha})\} \\ \text{subject to} \quad & \alpha_i \geq 0 \quad i = 1, \ldots, l \end{aligned} \tag{2.12}$$

なお，式 (2.12) では最小化問題の最適解では L の勾配が 0 になるため

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.13}$$

$$\vec{w} = \sum_{i=1}^{l} \alpha_i y_i \vec{x}_i \tag{2.14}$$

となる．したがって，式 (2.12) の問題は次のように書き換えることができる．

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j \\ \text{subject to} \quad & \alpha_i \geq 0 \quad i = 1, \ldots, l \\ & \sum_{i=1}^{l} \alpha_i y_i = 0 \end{aligned} \tag{2.15}$$

この双対問題を解く事により，SVM では識別に最適な識別超平面を導出することが可能である．

2.3 代表的な識別器の課題と解決策

前節の代表的な識別器において，それぞれに課題が存在する．ニューラルネットワークにおいて，学習器の入力データによってネットワークの重みの数が変化する．入力が高次元になる場合にはネットワークの構造が複雑になる為，誤差逆伝搬法による重みの更新演算において過学習が起こり，汎化能力が低下することが考えられる．また，SVM においてもマージン最小化の評価はサポートベクターとの距離のみが評価されるため，高次元のデータが入力された場合にいくつかの識別のパターンが考えられ，パラメータが決定される際に未知データに適さない識別超平面が選択される可能性がある．

これらの問題は入力空間が高次元になる場合に識別に有効な特徴量と識別に有効でない

特徴量が混在することでパラメータを一意に設定することが困難である事に起因する．これらの解決策として入力特徴ベクトルから識別に有効な特徴量のみを選択する変数選択法が考慮される．

変数選択のためには，すべての特徴の部分集合に対して予測性能を評価する必要がある．しかし，部分集合の数は特徴量数が増えると指数関数的に増大する．したがって，特徴量数が多い場合にすべての部分集合に対して評価することができない．そのため，比較的良い特徴量の部分集合を探策する手法が提案さている．単純な方法としては，Forward stepwise selection と Backward stepwise selection がある．Forward stepwise selection は特徴 1 個のみのモデルからはじめて特徴を 1 個づつ追加して行くことで，最も良い特徴の組を選び出す．逆に Backward stepwise selection は全ての特徴を含むモデルから特徴を 1 個づつ取り除いて行くことで，最も良い特徴の組を選び出す．しかし，高次元データの場合にはこれらの手法を用いた場合においても組み合わせの数が多く，最適な特徴の組を選択することが困難である．

そこで遺伝的アルゴリズム（GA）を用いた確率的な多点探索アルゴリズムによる特徴量選択法が提案されている[14) 15)]．複数の個体を用意し，特徴量を選択する組み合わせのパターンを遺伝的操作を用いて最適化することにより，最終的な特徴選択パターンを導出する手法である．近年，GA を構造的な遺伝子型の表現に拡張した遺伝的プログラミング（GP）が提案され様々な問題に適応されている．GP を用いることにより，特徴量選択よりも踏み込んだ特徴量変換を行うことが可能となった[16) 17) 18)]．本稿では GP による特徴量変換を用いて識別の精度向上に関する検討を行う．

3　遺伝的プログラミングを用いた特徴量変換の最適化

3.1　単目的 GP を用いた特徴量変換

前章で提起した識別する特徴量空間を低次元化する為の写像を GP を用いて最適化する．本章では，変換後の特徴量平面を単目的 GP を用いる場合について述べる．GP によって変換処理が行われた後に，それぞれの変換式が写像した後の特徴量平面が評価される．GP の評価には識別器により識別率を算出し，その識別率を評価値とする手法（wrapper approach）[19) 20)] と特徴量平面における情報エントロピーなどの情報量の純度を計る方法がある（filter approach）[21) 22)]．ここではより計算時間の短く，実装が単純である filter approach について最適化の流れを説明する．また，単目的 GP に対する識別に最適な特徴量変換法として，Neshatian らの既存手法では 1 次元の変換特徴量平面に対して，そのクラス分布が重なる領域を探索し．誤識別されたサンプルの個数を合計した情報量を算出を行っている[?)]．そこで述べられている最適化の流れについて説明する．

Step.1 Initialization

初期個体群として，ランダムに集団数だけの個体（木構造の変換式）を生成する．

Step.2 Evaluation

各個体の適合度を評価関数によって計算する．各個体による変換式によって特徴量を変換し，変換した特徴量においてクラス領域の分布の重なるデータの範囲を求める．この区間にあるデータの数を求め，これを評価値とする．遺伝的操作 (選択) の際に評価値が低い木構造程が高い評価となる．Fig.3 にデータの変換と評価値の概要図を示す．

Step.3 選択 (Selection)

集団の中から適合度を基準として， 次世代に残す個体を集団数だけ選択する．

Step.4 交叉 (Crossover)

集団の中からランダムに選択した 2 個体を対象とし，それぞれランダムに交叉点を選んだ後，枝を差し替え交配する．交叉の概要図について Fig.4 に示す．

Step.5 突然変異 (Mutation)

対象となる個体からランダムに突然変異点を選択し，突然変異点以降の木をランダムに作成した突然変異木と入れ替える．突然変異の概要図について Fig.4 に示す．

Step.6 終了条件 (Terminal Criterion)

予め決めておいた終了条件に到達するまで Step 2〜Step 5 を繰り返す．主な終了条件として探索世代数や目的の個体に達したか否かなどが挙げられる．Step.2 の評価手法では評価値が 0 となり，データの属性が完全に分離することを目的として最適化を行う．

3.2 偏りのある 2 クラスデータと識別率の評価

前節に示した流れにより，全体的な識別率の向上を図ることが可能である．しかし，このような単一の目的に対して各クラスのデータのサンプル数が異なる場合, この手法では大多数のサンプルを持つクラスの識別率が高くなり，一方のクラスの識別率が下がるように，識別が偏ってしまう傾向が見られる．Fig.5 にその一例を示す．この図では，本来であれば Line A の識別面が得られることが望ましいが，サンプル数に偏りがあるために，Line B の識別面が得られている．このような状況下では，識別率の信頼性は失われ，クラス間に差異のある傾向を見出した識別が行われない．

Bhowan らは，各クラスの識別率を算出してそのそれぞれの識別率を目的とした多目的最適化問題を解き，その結果として各クラスの識別率のバランスが良く，クラス間に差異のある識別を行うことのできる特徴量変換式を得ている[23]．本稿においても，これらの目的を取り扱う多目的問題最適化問題について検討する．

3.3 多目的最適化手法

多目的最適化手法とは，複数の評価基準を有する問題に対して適応され，複数の評価基準を同時に評価しながら最適化を行う手法である．多目的最適化問題の多くは，各々の評価基準がトレードオフの関係にあり，全ての評価基準において他の解よりも優れた解を導出することは困難であり，最適解をひとつに絞ることができない．そこで GA による多点探索により，最適解の候補となる解集合を導出する手法が考案されている．最適解の候補をパレート最適解と呼ぶ．パレート最適解とは「ある評価基準の値を改善する為には，少なくとも他の 1 つの評価基準を改悪せざるを得ないような解」の集合である．特に評価基準として用いられる目的関数を 2 目的とした場合のパレート最適解を Fig.6 に示す．本稿では，多目的 GA のアプローチのうちの NSGA-II[24] と SPEA-II[25] を用いた．変換式の最適化についての定式化については次節で述べる．

3.3.1 NSGA-II

NSGA-II は Deb, Agrawal によって提案された多目的遺伝的アルゴリズムのひとつである．NSGA-II は GA の遺伝的処理によって得られた解に対して，ランク付け・混雑度トーナメント選択・非優越ソートを用いることによってパレート解の探索を行う．NSGA-II では，保存する母集団 P_t と交叉・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母集団 Q_t の 2 つの独立した母集団を用いて解探索を進めていく．

具体的には，まず世代 t における親母集団 P_t から遺伝的操作を用いた探索を行うための子母集団 Q_t を選択する．Q_t に対して各遺伝的操作を行い Q_t を更新する．次に，各遺伝的操作を行った Q_t と親母集団 P_t を組み合わせた $R_t = P_t \cup Q_t$ を生成し，選択操作によって個体数 $2N$ の R_t から個体数 N の P_{t+1} を新たに選択し探索を進めていく．NSGA-II の流れを以下に示す．

step.1 親母集団と子母集団を組み合わせて $R_t = P_t \cup Q_t$ を生成する．R_t に対して非優越

ソートを行い，全個体をフロント毎 (ランク毎) に分類する：$Fi, i = 1, 2, ..., etc.$ 個体のランクの付け方について Fig.8 に示す．非優劣ソートでは個体群の中から非劣個体を求めランクを付加し，その後，非劣個体の消去とランク値の更新を行っていく．この操作によって，全ての個体にランクが付加される．

step.2 新たな母集団 P_{t+1} を生成．変数 $i = 1$ とする．

$|P_{t+1}| + |F_i| < N$ を満たすまで，$P_{t+1} = P_{t+1} \cup F_i$ と $i = i + 1$ を実行する．

step.3 混雑度ソート (Crowding-sort) を実行し，最も多様に広がっていた個体 $N - |P_{t+1}|$ 個を P_{t+1} に加える．混雑距離の概要図を Fig.9 に示す．混雑距離 (Crowding Distance) とは，ある個体 i の周りにある個体の密度を評価するための手法である．混雑距離は，同一ランク (同一フロント内) 同士で用いられ，各目的関数軸において隣り合う個体間との距離を足し合わせたものである．各目的関数の最大値と最小値をとる個体には最大距離，もしくは無限距離を与える．それに対して境界個体以外の全ての個体 $(j = 2, ..., l - 1)$ に対しては式 (3.1) に従った混雑度計算を行う．

$$d_{I_j^m} = d_{I_j^m} + \frac{f_m^{I_{j+1}^m} - f_m^{I_{j-1}^m}}{f_m^{\max} - f_m^{\min}} \tag{3.1}$$

step.4 P_{t+1} を基に，混雑度トーナメント選択，交叉，突然変異を用いて新たな子母集団 Q_{t+1} を生成する．

このように NSGA-II では，親母集団 P_t と子母集団 Q_t を組み合わせた母集団 R_t の上位 N 個体を選択し，次世代の親個体 P_{t+1} としている．また，探索個体 (子個体)Q_t は，親個体 P_t から混雑度トーナメント選択を用いて選抜されており，親個体 P_t のより優れた個体を用いて各遺伝的操作を用いた探索が行われている．一方，常に優良個体を保存する親母集団 P_t と探索を行う子母集団 Q_t を分けて保持することにより，それまでの探索で発見した優れた解が欠落するのを防いでいる．NSGA-II の親母集団 P_t 更新の概念図を Fig.7 に示す．

3.3.2 SPEA-II

SPEA-II は SPEA の改良アルゴリズムとして Zitler らが 2001 年に提案した多目的 GA 手法である．SPEA-II は GA の遺伝的操作によって得られた解に対して各個体の支配個体数と被支配個体数による適合度評価，並びにアーカイブ個体に優良個体を常に保持し，その個体群から探索個体群を提供する機構とアーカイブ個体を適切な個体数に削減する機構を併せ持つ．SPEA-II による探索は NSGA-II と類似する部分があり，アーカイブ個体群と探索個体群という 2 つの個体群を用いて探索を進める点は共通である．SPEA-II のアルゴリズムの流れを以下に示す．

step.1 初期化：初期母集団 Q_0 を生成する．空のアーカイブを生成する：P_0=0,Set t = 0

step.2 適合度割当て：R_t=Q_t+P_t とし，R_t における個体の適合度値を計算する．適合度割当てでは，Fig.10 に示したように，まず全ての個体 i に対して支配している個体の数 $s(i)$ を求める．次に，各個体を支配している個体の持つ $s(i)$ を全て足し合わせた値を，その個体の適合度値 $f(i)$ とする．最終的な適合度値算出の概要図について Fig.11 に示す．

step.3 環境選択：R_t における全ての非劣個体を P_{t+1} へコピーする．$|P_{t+1}|$= N の場合は環境選択を終了する．$|P_{t+1}| >$ N の場合には，Fig.12 に示すように端切りオペレータを用いて $|P_{t+1}|$ の個体数を N に削減する．また，$|P_{t+1}| <$ N の場合には，Qt における優良個体 $N - |P_{t+1}|$ 個分を P_{t+1} へコピーし，P_{t+1} の個体数を N にする．

step.4 終了判定：もし t ＜終了世代数，もしくはその他の終了条件が満たされた場合，P_{t+1} の中の非劣個体群が最終的な解として出力され，探索は終了する．そうでなければ Step 5 に進む．

step.5 メイティング選択：P_{t+1} からバイナリトーナメント選択によって N 個分の Q_{t+1} を選択する．

step.6 交叉および突然変異：Q_{t+1} に対して交叉と突然変異オペレータを実行する．Step 2 から Step 6 を繰り返す．

このように，SPEA2 では 2 つの母集団（アーカイブ集団 P，探索母集団 Q）を用いて探索を進める点で NSGA-II と共通である．

3.4 多目的 GP を用いた特徴量変換

本稿では，多目的最適化の目的関数には各クラスの識別率である真陽性率（True Positive Rate：TPR）と真陰性率（True Negative Rate：TNR）を用いている．これらは一般的にトレードオフの関係にある．Fig.13 に 2 クラス識別における識別における要素を示し，Eq.3.2 と Eq.3.3 に TPR と TNR の導出式を示す．

$$TPR = \frac{TP}{TP + FN} \tag{3.2}$$

$$TNR = \frac{TN}{TN + FP} \tag{3.3}$$

多目的最適化手法には，代表的な NSGA-II[24] と SPEA-II[25] を利用した．近年では，MOEA/D[26] や NSGA3[27] といった強力な手法も提案されているが，われわれの手法を多目的に適用した場合に，既存の研究[23] と比較するために，本稿ではこれらの手法を選択している．

3.5　識別のための閾値の決定

GP の変換式によって変換された 1 次元特徴量平面における識別手法について述べる．既存手法では，変換後の特徴量値を正負で識別した[23]．すなわち，識別のための閾値は 0 である．それに対して本稿では，次の 2 つの閾値設定手法を検討する．1 つ目の閾値設定手法は，1 次元へのデータ変換後にすべての候補となる閾値に対して学習データすべてを考慮して最も識別率の高い閾値を考慮する方法である．一見，最適な方法と考えられるが，学習データのクラス間データ数の偏りによる学習の偏りによって未学習データに対してはうまく識別できない可能性がある．もう一方の手法は各クラスのデータ数の比に応じて，全体のデータのどの部分に識別の閾値を設定すればよいかを決定する方法である．

3.5.1　閾値の候補の総当たり法

1 次元に変換された学習データのサンプル間を閾値の候補とする．これらの候補となる閾値に対して，最も学習データの識別率が良い値を，最終的な閾値に採用する方法である．候補となる閾値の全探索手法を用いる（以下，閾値総当たり法：all）．それぞれの学習データの間隔に設定する閾値の設定方法は，閾値の両側のサンプルの平均値とする．

3.5.2　学習データ数の比から閾値を設定する手法

本稿では多次元データは 1 次元データに変換される．学習データにおける 2 クラスに属するデータ数がそれぞれ $n1$ および $n2$ である場合には，$n1 : n2$ に識別するような閾値が設定されることが理想である．$n1 : n2$ に識別する閾値を検討する（以下，データ数比法：ideal）．境界値の導出方法は，事前にサンプル数に応じた理想的なクラス分布から境界の両側のサンプルの順位を求める．そして，学習データの 1 次元変換値において，境界両側の順位に当たるサンプルを平均して識別境界値を得る．Fig.14 に概要図を示す．データ数に偏りのある場合にこの境界値を目標とした最適化は有効と考えられる．

4　多目的遺伝的プログラミングを用いた特徴量変換についての検討

4.1　評価実験概要

多目的最適化手法の NSGA-II と SPEA-II を用いて TPR と TNR の 2 つの評価値を目的とした多目的最適化問題を解く．識別率の算出する際の閾値の決定手法として閾値総当たり法とデータ数比法，並びに既存手法について検討を行う．

4.2　実験方法

実験データに対しては，2 つのデータを用いる．大気圏の層の分類についての Ionosphere データ[28]（データ数 350，2 クラスデータ比 1:2），心疾患についての Spect Heeart データ（データ数 256，2 クラスデータ比 1:4）の 2 種類のデータを用いる．前者のデータに対して，後者のデータは各クラスの偏りが大きいデータである．また，識別率の検証は既存手法の条件と合致させるため，2fold-crossvalidation を用いる．NSGA-II と SPEA-II の各手法の 2 つの閾値設定手法のパレート解を導出し比較検討を行う．また，単目的 GP によって導出された解，SVM の線形カーネルによって得られた解と既存手法のパレート解による比較も同時に行う．また，表.1 に今回の実験で用いる MOGP のパラメータを示す．このパラメータについても既存手法と同様のものを用いた．

4.3　実験結果

4.3.1　各最適化手法におけるパレート解の導出結果

NSGA-II と SPEA-II を用いた際の 2 つの識別手法（all，ideal）のそれぞれにおいて，10 試行で最大の識別率を持つ個体を含んだパレート解を導出し，TNR，TPR の 2 目的の評価値空間に図示した．また，その結果を Fig.15 と Fig.16 に示す．Fig.15 は Ionosophere データについてのパレート解の導出結果であり，Fig.16 は Spect データについてのパレート解の導出結果である．Fig.15 と Fig.16 に単目的 GP，baseline（SVM）と既存手法（Bhowan らの手法）によって得たパレート解も併せて図示した．Fig.15 において，Ionosophere データではパレート中央付近の解において，既存手法よりも検討手法の精度が上回っている．2 つの最適化手法でのパレート解の精度の差異は見られないが，データ数比法は閾値総当たり法とは異なり，左上の領域の解が見られなかった．データ数の偏りの少ないデータにおいて，データ数比法では学習データの分布に対して一定の値を閾値に用いる為，多様性の維持が困難であることが考えられる．一方，Fig.16 において，Spect データでは NSGA-II と SPEA-II の 2 つの識別手法でデータ数比法が閾値総当たり法をパレート解の識別精度で上回った．特にパレート解の中央付近の解に差が表れている．これらの検討手法は既存手法によって得たパレート解の精度を特にパレートの右下の領域で上回っている．また，Fig.15 と Fig.16 の 2 つの単目的 GP と baseline の結果に比べ，多目的最適化手法によって得たパ

レート解の識別精度は TNR と TPR を総合して上回っている．

4.3.2　各最適化手法における最大識別率の比較

TPR と TNR の最大化に関して，特に NSGA-II（ideal) を Spect データに対して行った際のパレート最適フロントの性質について示す．実験．4.3.1 の Spect データにおいて，NSGA-II（ideal）に対するパレート最適フロントから TPR が高い解から順に high，middle，low の 3 つの個体を抽出し，その個体の持つ特徴量平面について調べた．Fig.17，Fig.18，Fig.19 に特徴量平面のヒストグラムを示す．TPR が高い領域では positive クラスの分布が一方の領域に集まっている．TPR が低下するにつれ positive クラスの分布は広がるが，一方で，negative クラスの分布がもう一方の領域に集まるようになる．Fig.19 のようなデータ数の多い negative クラスの識別率 TNR が高いときに全体的な識別率は最も高くなるが，Fig.18 のように両クラスの識別率が高い時に，識別線に対して各クラスのデータ頻度のバランスの良い分布が得られている．また，医療データにおいては悪性のデータ数を集めることが難しく，またその識別率が重要になる場合があり，TPR が重要となる．その場合には Fig.17 のような特徴量平面が識別に有効である．

4.3.3　NSGA-II と SPEA-II の比較

本稿では NSGA-II と SPEA-II の二つの最適化手法を用いた．今回の実験結果において，各々の最適化手法において目立った差異は見られなかった．その理由として，両手法は 2 つの母集団を用いて探索を進める点やエリート個体を残す機構を有する点，さらには多様性を維持する機構において類似していることが考えられる．それぞれの機構のアルゴリズムの差異は存在するが，今回の課題においてはその影響はほとんど見られなかった．他の最適化手法を用いた時の影響について考える必要がある．

4.3.4　データ数比法と閾値総当たり法の比較

パレート解の導出結果により，データ数の偏りが 1:2 である Ionosophere データにおいて，識別手法としては閾値総当たり法とデータ数比法の間に差異は見られなかった．この結果はデータ数の偏りが比較的少ないことにより，学習に識別率のバランスの影響がでなかったことによると考えられる．一方，データ数の偏りが 1:4 である Spect データにおいて，識別手法としてデータ数比法が閾値総当たり法を上回った．これについて，Fig.18 と Fig.21 に各識別手法の 1 次元変換特徴量のヒストグラムを示す．また，これらの解の評価値空間での散布図を Fig.20 に示す．比較する 2 つの特徴量平面はパレート中央付近の解である．Fig.21 の結果より，閾値総当たり法では識別率としては良い解が算出されたと言えるが，1 次元変換特徴量平面では識別線よりも遠くに誤識別されたサンプルの分布を確認することができる．それに対し，Fig.21 の結果により，データ数比法によって導出された最も識別精度の高い解における 1 次元変換特徴量平面では誤識別されたサンプルが識別線付近にある．これにより，クラス間のデータ数に偏りのある場合，データ数比法は閾値総当たり法よりも識別に対してパレート解中央付近の解において優れており，結果として信頼性の高い変換特徴量平面を得ることが示唆された．

5　結論

本論文では，多目的遺伝的プログラミング（多目的 GP）を用い，高次元データから識別に有効な 1 次元の特徴量に変換し，識別を多目的最適化するアルゴリズムの検討を行った．その際，1 次元変換特徴量平面における識別手法や多目的最適化手法の比較検討を行った．

本アルゴリズムで構築される変換式は，簡易な演算子を木構造状に組み合わせたものである．構築した木構造の変換式によって変換される特徴量の評価を単一の目的で行った場合，クラス間のサンプル数に偏りが存在する為に，大多数のサンプルを内包するクラスに識別率が依ってしまう傾向が見られた．そのため，多目的最適化によって，2 クラスの識別に対してバランスの良い特徴量変換式の探索を行った．多目的最適化の評価指標には，真陽性率（TPR）と真陰性率（TNR）を用いた．

検討した多目的最適化手法には，NSGAII と SPEAII を適応し，パレート最適フロントの比較検討を行った．識別手法には学習データにおける 1 次元特徴量平面における識別閾値を総当たりで決定する閾値総当たり法と，学習データ数のバランスを考慮して，理想的な閾値を抽出するデータ数比法の 2 つの手法についてパレート最適フロントの比較及び 1 次元変換特徴量平面における特徴量分布の比較検討を行った．また，既存手法とのパレート最適フロントの精度の比較も併せて行った．

実験の結果として，最適化手法において NSGAII と SPEAII の間にパレート解の精度として，有為な差異は見られなかった．識別手法において，既存手法の識別手法よりも検討した 2 つの手法がパレート最適フロントにおいて優れた解を導出した．また，閾値総当たり法とデータ数比法においては，クラス間サンプル数の偏りの大きいデータにおいおて，データ数比法が良好なパレート最適フロントを導出した．次に，変換された 1 次元特徴量平面においても，データ数比法は閾値総当たり法に対して誤識別のサンプルが識別線の付近にあり，理想的な 2 クラス分布により近くなっている．これにより，データ数比法は識別に有効な 1 次元特徴量平面を構築することが示された．以上の結果より多目的最適化手法の単目的最適化手法に対する有効性が示唆され，また，識別手法が解の精度に影響を与えることが示唆された．

謝辞

本研究を遂行するにあたり，熱心なご指導，多くのご協力を頂きました，同志社大学生命医科学部の廣安知之教授に心より感謝致します．廣安教授の下で研究することができ，日々のミーティングにおいてご指摘を受け議論を深めることによって，研究を深めることができました．結果として学会に参加し議論を行うことができ，非常に多くのことを学ぶことができました．また，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山本詩子助教に心より感謝致します．私の指導院生である吉田倫也氏には，研究面や様々な面で協力頂き心より感謝しております，いつも前向きに研究をおこなうことができたのは吉田氏の心強い支えのおかげでした．そして私の所属するデータマイニング班の塙賢哉君，佐藤琢磨君，田村陵大君，勝林洋介君とは多くの議論を共に行ってきました．貴重な経験を頂き誠に感謝しております．最後に，医療情報システム研究室の皆様の御陰で，私はこの二年間充実した研究生活を送る事ができました．この場を借りて厚く御礼申し上げます．ありがとうございました．

参考文献

1) C.M. BISHOP, 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.

2) C.C. Bojarczuk and H.S. Lopes and A.A. Freitas and E.L. Michalkiewicz, A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets, *Artificial Intelligence in Medicine*, vol 30, 2004, pp. 27-48.

3) A. Moreda-Pieiro, A. Fisher, S.J. Hill, The classification of tea according to region of origin using pattern recognition techniques and trace metal data, *Journal of Food Composition and Analysis, Elseveir*, April 2003, pp 195-211.

4) B. Wray, Learning classification trees, *Statistics and Computing*, vol 2, June 1992, pp. 63-73.

5) G.E. Hinton and R.R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, *Science*, Vol. 313, No. 5786, 2006, pp. 504-507.

6) V.N. Vapnik, Statistical Learning Theory, Wiley, 1998.

7) F. Otero, M. Silva, A. Freitas, and J. Nievola, Genetic programming for attribute construction in data mining, in Genetic Programming (Lecture Notes in Computer Science, vol. 2610.) Berlin, Germany: Springer, 2003, pp. 101-121.

8) J. Koza:Genetic programming,on the programming of computers by means of natural selection, MIT Press, 1992.

9) O. Smart, H. Firpi, and G. Vachtsevanos, “Genetic programming of conventional features to detect seizure precursors,” *Eng. Applicat. Artif. Intell.*, vol. 20, no. 8, 2007, pp. 1070-1085.

10) H. Guo and A.K. Nandi, “Breast cancer diagnosis using genetic programming generated feature,” *Pattern Recognit.*, vol. 39, no. 5, May 2006, pp. 980-987.

11) G. Hong, L.B. Jack, and A.K. Nandi,“Automated feature extraction using genetic programming for bearing condition monitoring,” in *Proc.14th IEEE Signal Process. Soc. Workshop*, Sep.-Oct. 2004, pp. 519-528.

12) V. 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, 1998, pp.55-85.

13) V. Vapnik, ”Statistical learning theory”, John Wiley, New York, 1998.

14) J.H. Yang, V. Honavar, Feature subset selection using a genetic algorithm, *IEEE Intelligent Systems*, 1998, pp. 44-49.

15) Z. Zhu, Y.S. Ong, M. Dash, Markov blanket-embedded genetic algorithm for gene selection, Pattern Recognition, 2007, pp. 3236-3248.

16) M.G. Smith and L. Bull, “Genetic programming with a genetic algorithm for feature construction and selection,” *Genet. Programming Evolvable Mach.*, vol. 6, no. 3, pp. 265-281, 2005.

17) H. Guo, Q. Zhang, and A.K. Nandi, “Feature extraction and dimension- ality reduction by genetic programming based on the Fisher criterion,” *Expert Syst.*, vol. 25, no. 5, pp. 444-459, 2008.

18) R. Kohavi and G. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1-2, pp. 273-324, Dec. 1997.

19) M. Muharram and G.D. Smith, "*Evolutionary constructive induction,*" *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 11, pp. 1518-1528, Nov. 2005.

20) M.G. Smith and L. Bull, "Genetic programming with a genetic algorithm for feature construction and selection," *Genet. Programming Evolvable Mach.*, vol. 6, no. 3, pp. 265-281, 2005.

21) D.J.C. MacKay, *Information Theory, Inference, and Learning Algo- rithms.* Cambridge, MA: Cambridge Univ. Press, Oct. 2003.

22) J.C. Principe, D. Xu, and J. Fisher, "Information theoretic learning," *in Unsupervised Adaptive Filtering.* New York: Wiley, 2000, pp. 265-319.

23) U. Bhowan, M. Johnston, M. Zhang, and X. Yao, Evolving Diverse Ensembles Using Genetic Programming for Classification With Unbalanced Data, *IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION* , VOL. 17, NO. 3, June 2013.

24) K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast elitist multiobjective genetic algorithm: NSGA-II", *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, Apr. 2002, pp. 182-197.

25) E. Zitzler, M. Laumanns, and L. Thiele, "Spea2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization", *Dept. Electr. Eng.*, Swiss Federal Instit. Technol., Zurich, Switzerland, TIK Rep. 103, 2001.

26) Q. Zhang and H. Li, "MOEA/D: A multi-objective evolutionary algorithm based on decomposition", *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, Dec. 2007, pp. 712-731.

27) K. Deb and H. Jain, "An improved NSGA-II procedure for manyobjective optimization, Part I: Problems with box constraints", *IEEE Trans. Evol. Comput.*, pp. 1-8.

28) D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz , UCI repository of Machine Learning databases, 1998. Available: `http://archive.ics.uci.edu/ml/datasets.html`

付 図

付 表

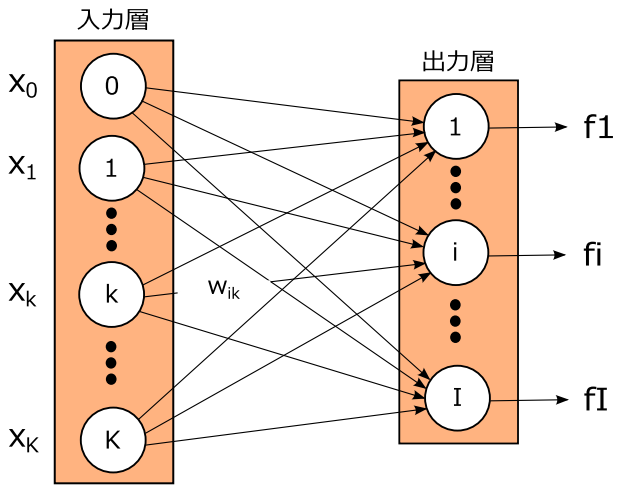

Fig. 1　単層パーセプトロン

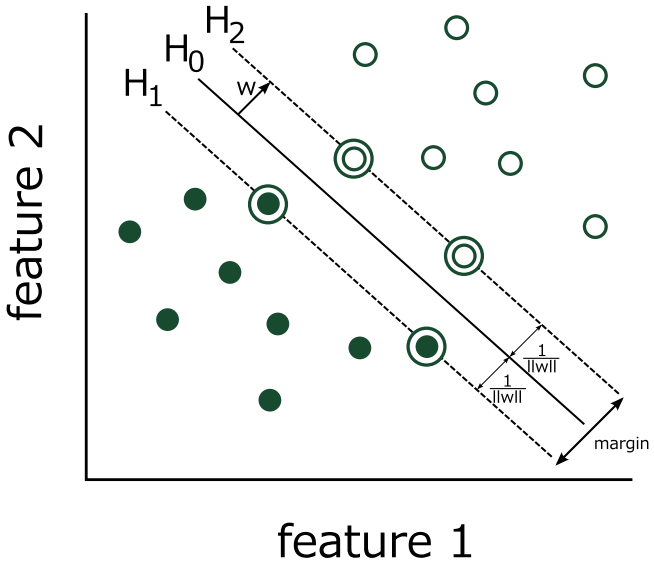

Fig. 2　ハードマージン SVM

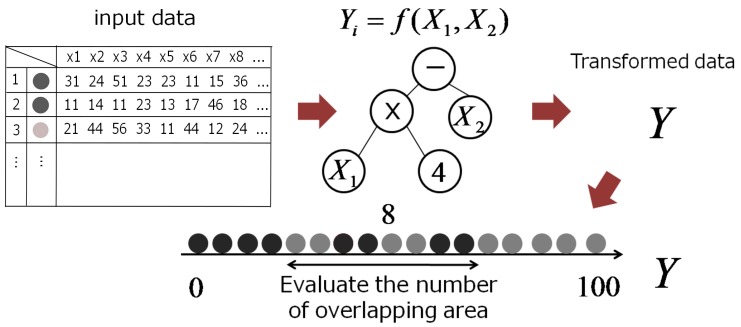

Fig. 3　変換式の評価

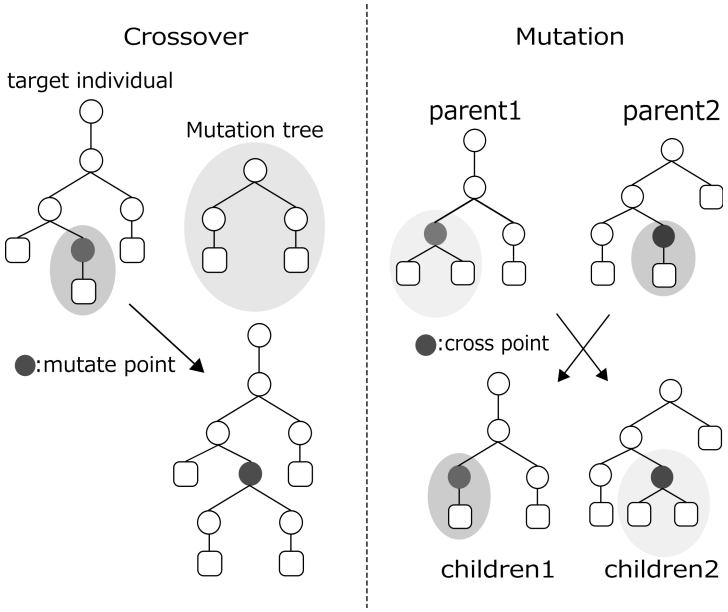

Fig. 4　交叉・突然変異

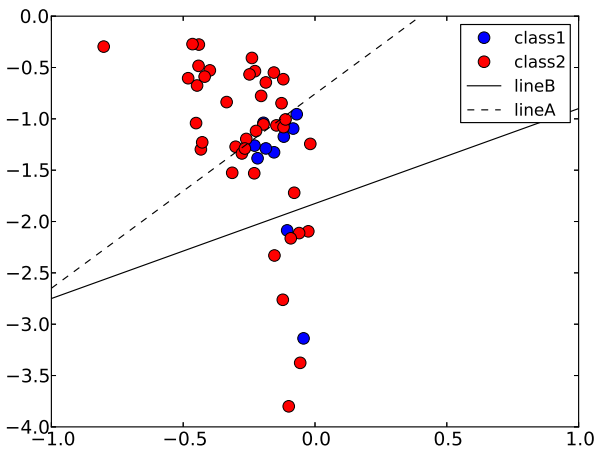

Fig. 5　信頼性の低い識別の例

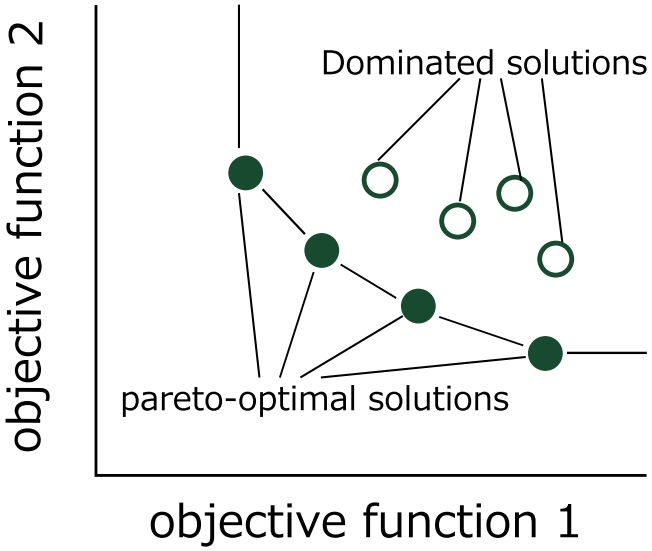

Fig. 6　2 目的最小化問題におけるパレート最適解

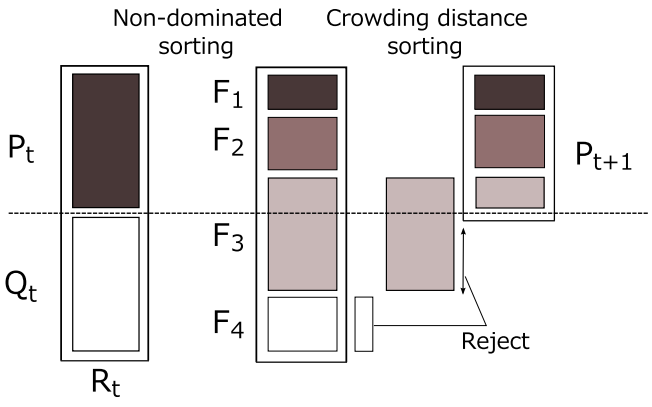

Fig. 7　NSGAII の母集団更新方法

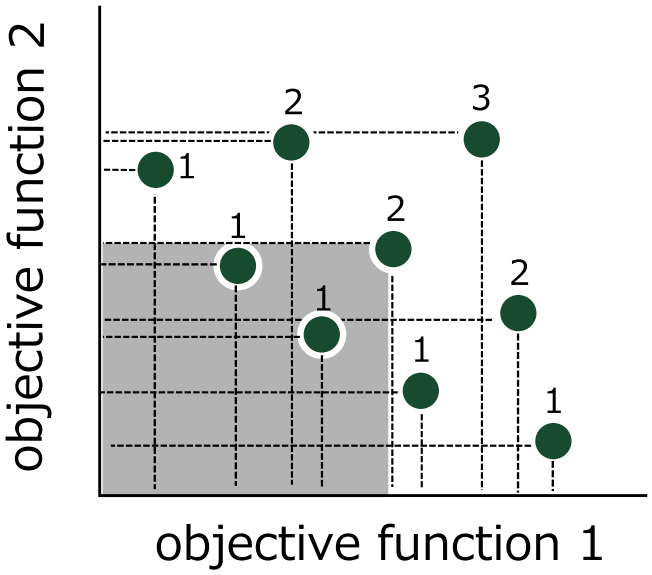

Fig. 8　非優越ソートを用いたランク付け

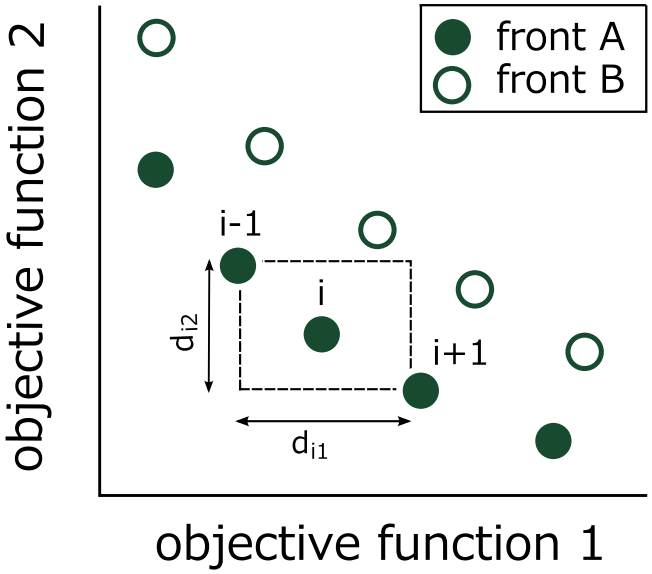

Fig. 9　混雑距離の計算

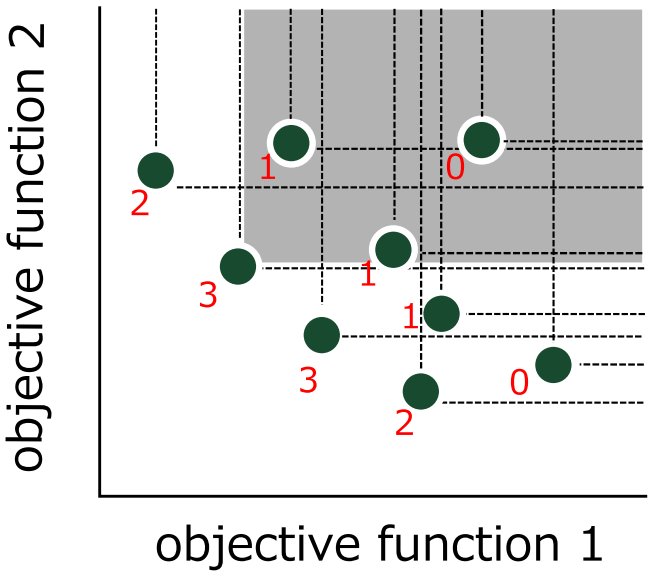

Fig. 10　各個体の優越個体数の算出

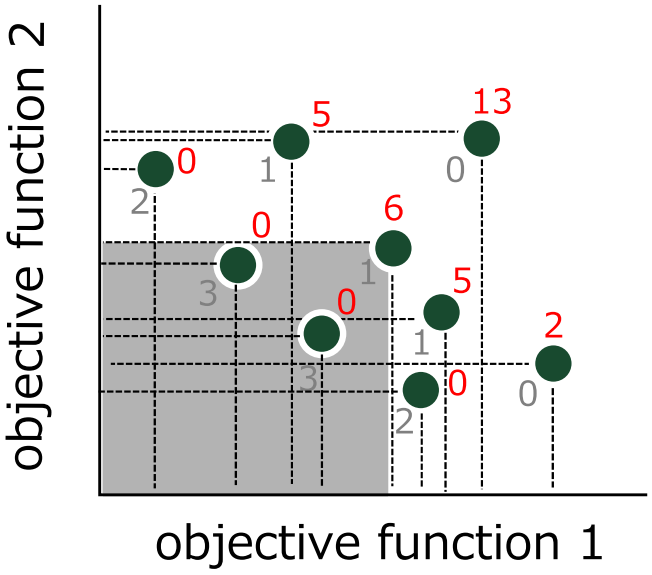

Fig. 11　SPEAII の評価値の算出

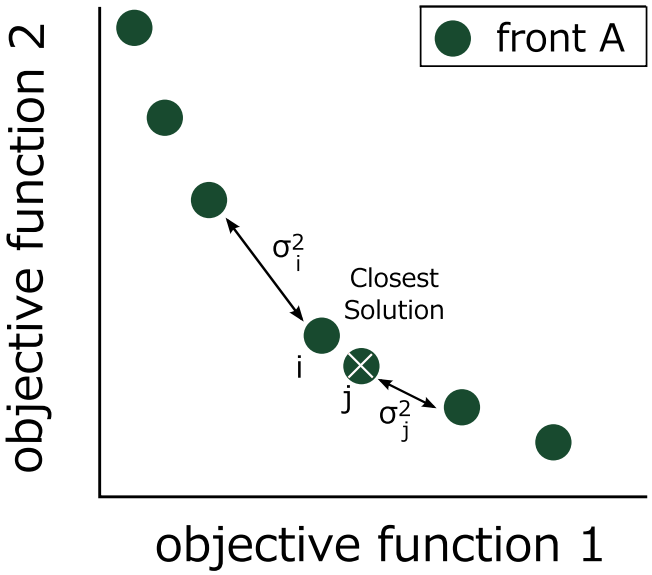

Fig. 12　アーカイブ端切り法

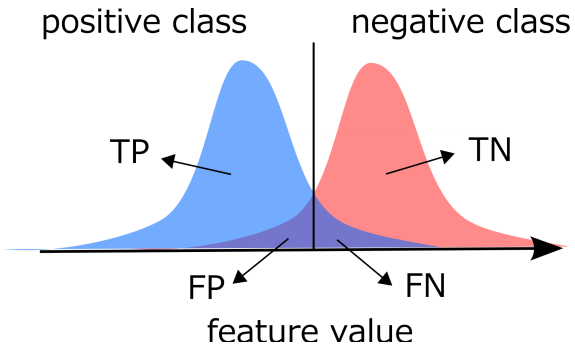

Fig. 13　2 クラス識別概要図

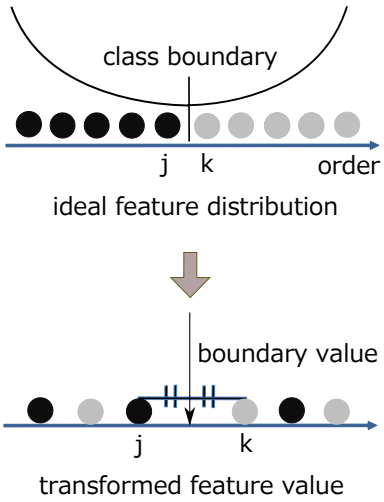

Fig. 14　識別境界値

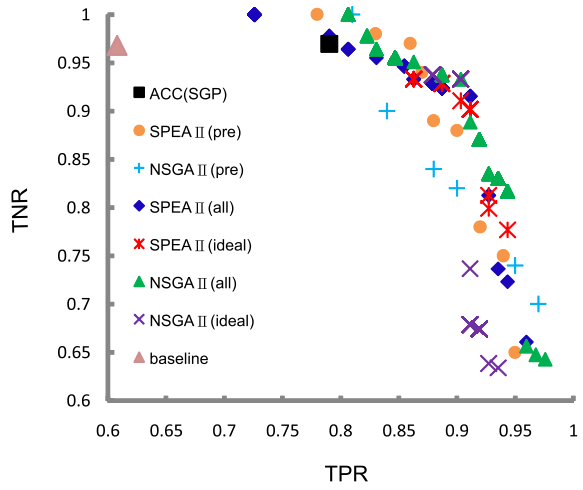

Fig. 15　パレート解の導出結果（Ionosophere）

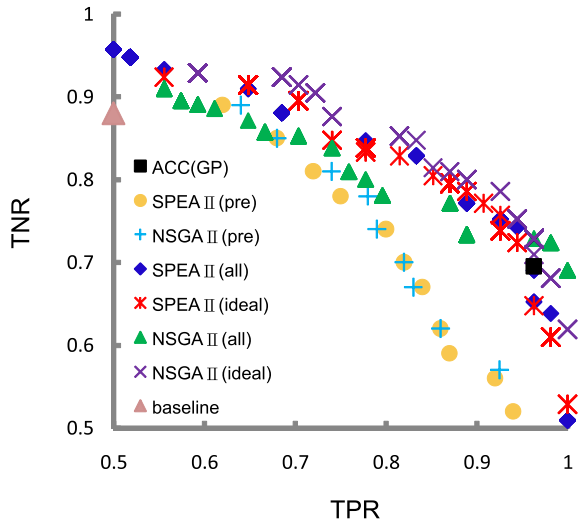

Fig. 16　パレート解の導出結果（Spect Heart）

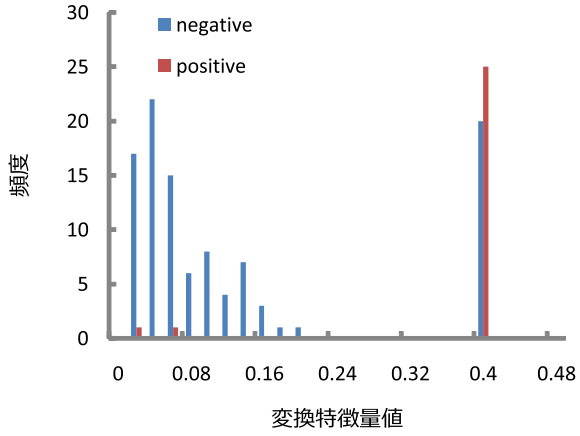

Fig. 17　特徴量平面（ideal:high）

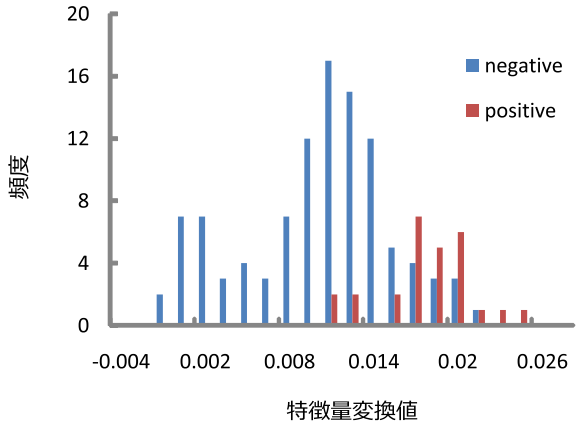

Fig. 18　特徴量平面（ideal:middle）

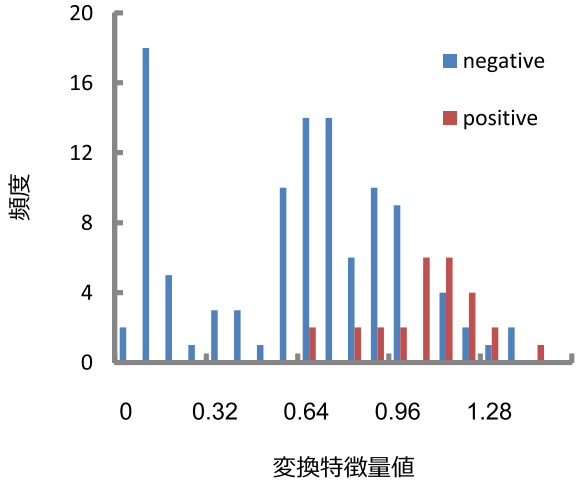

Fig. 19　特徴量平面（ideal:low）

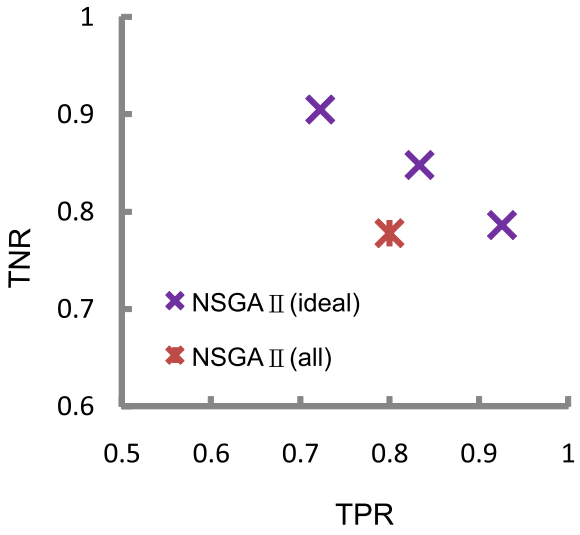

Fig. 20　抽出した解の評価値空間

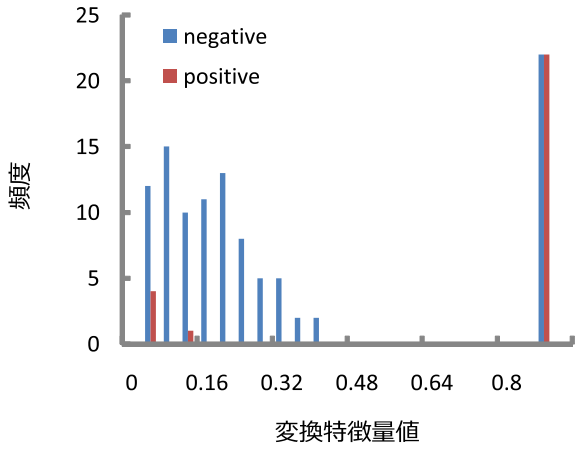

Fig. 21　特徴量平面（all:middle）

Table 1　多目的 GP 及び GP で用いたパラメータ

パラメータ	値
個体数	500
世代数	50
交叉率	0.6
突然変異率	0.4
選択方法	トーナメント選択
トーナメント数	2
木の深さ制限	8
関数ノードの種類	＋，－，×，÷
ペナルティ	10^{-3}
試行回数	10

Table 2　baseline で用いる SVM のパラメータ

SVM パラメータ	値
Cost parameter	1
Kernel	線形カーネル
Dimension	1