原著論文

pdi-Baggingの定式化とその評価†

入江　穂乃香*1・林　勲*1

パターン分類問題に対して，複数個の弱判別器を学習しそれらを統合的に組み合わせて全体の識別精度を向上させるアンサンブル法に対する関心が高まっている．本論文では，仮想的に生成したデータを学習データに追加して，複数個の弱判別器を学習しこれらの多数決により識別精度を向上させる新たなバギング (Bagging) を提案する．仮想的に生成したデータをバーチャルデータと呼び，このバギングを pdi-Bagging(Possibilistic Data Interpolation-Bagging) と呼ぶ．ここでは，バーチャルデータの領域を特定化する新たなバーチャルデータ発生法とバーチャルデータのクラスを決定する新たな評価指標を提案する．バーチャルデータの発生領域の特定化では，学習データの分布性と方位性を考慮し，誤判別型，正判別型，混合判別型の 5 種類の発生法を導入する．また，評価指標の定式化では，正誤判別データとバーチャルデータとの空間的類似指標を基盤とした 3 種類の評価指標を定式化する．pdi-Bagging では，バーチャルデータが学習データに追加されるので，バーチャルデータの追加の方法によってはクラス間のデータ量の偏りがなくなり，判別線の同定精度が向上し，評価データに対して高い識別率を得ることができる．ここでは，新たな pdi-Bagging のアルゴリズムを定式化し，数値例を用いて本手法の有用性を議論する．

キーワード：ファジィ推論，バーチャルデータ，アンサンブル法，バギング，クラスタリング

1. はじめに

パターン分類問題に対して，複数個の弱判別器を統合的に組み合わせて全体の識別精度を向上させるアンサンブル法 [1–5] に対する関心が高まっている．アンサンブル法は，学習データを用いて複数の弱判別器を学習し，評価データに対して，それらの弱判別器を結合して出力を得る．アンサンブル法は，弱判別器を簡便に結合する判別器結合モデルとクラスパターンに高い相関をもつ属性で弱判別器を構成する属性結合モデルに分類できる [1]．属性結合モデルは広義のアンサンブル法であり，混在データ利用法や最大エントロピー法などがある．一方，判別器結合モデルは，それぞれの判別器が独立的に結合する独立型と依存関係を保持しながら結合する依存型に分類できる．

判別器結合モデルの独立型は，サンプリングによってデータ集合から複数の学習データを構成し，それぞれの判別器が個別の学習データによって学習され，それらを独立的に統合して高い識別率を得る手法である．バギング [6] やランダムフォレスト [7, 8]，誤り訂正符号法 [9] がこの分類に属する．バギングは，サンプリングされた学習データに対して複数の判別器が独立に学習され，それらの判別器の多数決もしくは判別関数の平均によって評価データに対する識別を求める手法である．バギングは，アルゴリズム構造がシンプルであるので，医療データのクラスタリングモデル [6] や時系列の予測モデル [10] として多用されている．また，半導体ウェハの欠陥を検出するモデルとして AdaBoost よりも高精度を得た事例もあり [11]，SSM 調査 (社会階層と社会移動に関する全国調査) に適用した事例もある [12]．ランダムフォレストは，パターン識別や回帰分析，クラスラリング法，画像処理手法のアルゴリズムとして多用されており，Open CV，MATLAB，FORTRAN，Waffles，R，Microsoft の Kinect for Xbox 360 にもライブラリが実装されている [7, 13].

一方，判別器結合モデルの依存型とは，サンプリングによって複数の学習データが構成された場合，複数の判別器が逐次的もしくは同時に依存関係を保ちながら学習され，最終的にはそれらを統合して高い識別率を得る手法である．逐次的タイプとしてブースティング (Boosting) [14] があり，同時学習タイプとして混合エキスパート法 [15] がある．混合エキスパート法は，ゲートネットワークを用いて複数の弱判別器の評価を結合し，それぞれの弱判別器の中で最大値を出力する弱判別器の結果を最終結果とする手法である．混合エキスパート法によるブログ検索等への応用の報告がある [16]．ブースティングは，逐次的に弱判別器を学習させ，識別率を向上させる手法であり，AdaBoost [14, 17] は特に有用で，データ集合の特徴量を解析しやすい利点がある．人や車両等の画像認識に多く用いられている [18]．また，汎化精度を向上させるため，学習データの分割法 [19] の提案や 2 クラス識別器の Gentle Boosting [20] をマルチクラスに拡張した Joint Boosting [21] なども提案されている．特に，Joint Boosting はマルチクラス Boosting [22] であるため，画像処理における多クラス判別に多用され，コンビニやスーパーマーケット等での購買者の行動検出のための画像認識 [23] やカメラセンサによる人物トラッキングの精度向上 [24] 等に多用されている．このように，ブースティングに代表される依存型は，学習データに対して複数の弱判別器が逐次的な相互依存関係を保ちながら学習され，その依存関係を有した入出力関係を同定できる．一方，バギングに代表される独立型は，それぞれの学習データは個別に構成されて独立であるが，処理アルゴリズムは比較的簡便であり，精度が高いという

† Formulation of pdi-Bagging and Its Evaluation
Honoka IRIE and Isao HAYASHI

*1 関西大学大学院 総合情報学研究科
Graduate School of Informatics, Kansai University

特徴をもつ．なお，バギングとブースティングを統合したアンサンブル法も提案されている [25].

我々は，弱判別器の学習時に仮想的にデータを生成し学習データに追加して，判別線の識別率を向上させるバギングを提案している．このバギングを pdi-Bagging(Possibilistic Data Interpolation-Bagging) と呼び，仮想的に生成したデータをバーチャルデータと呼ぶ [26, 27]．バーチャルデータの追加で学習時にデータ量が増えるので，クラス間でデータ量の偏りをなくすことが可能で，各クラスのデータ量が均等となり，判別線の同定精度が向上する．しかし，バーチャルデータを発生させる発生元の学習データの正誤判別の種類とその発生領域の特定化，及びバーチャルデータのクラス付与法によって全体の認識精度が異なってくる．

本論文では，バーチャルデータ発生元の学習データの種類と発生領域を特定化し，さらに，バーチャルデータのクラス付与法を定式化する新たなバギング (pdi-Bagging) を提案する [28, 29]．具体的には，発生領域の特定化では，バーチャルデータ発生元の学習データのクラスが誤判別と識別された場合や正判別と判別された場合によって発生方法を変更するバギングを定式化する．発生元の学習データのクラスが誤判別と識別された場合の誤判別型では，判別線の領域に集中してバーチャルデータを発生させる．誤判別クラスのデータは判別線に近い位置に存在するので，誤判別データ周辺で発生したバーチャルデータも判別線近傍に分布する．発生元の学習データのクラスが正判別と識別された正判別型では，バーチャルデータは全データ空間に均一に発生させるか，誤判別データと正判別データの分布性と方位性を考慮して特定領域に発生させる．正判別クラスの学習データは判別線の近傍に位置するとは限らないので，発生したバーチャルデータは全データ空間に均一に分布するか，特定化された領域に集中して分布することになる．さらに，これらの正判別と誤判別を混合させた混合型のバギングのアルゴリズムも定式化する．ここでは，pdi-Bagging として，5 種類のバーチャルデータの発生法を定式化し，その有用性を議論する，一方，バーチャルデータの教師クラスを変更するための評価指標として，多次元上でのユークリッド距離を基に正誤判別データとの類似度を導入した新たな評価式を定式化する．評価式は，バーチャルデータの正誤判別データからの距離，クラスの中心からの距離，バーチャルデータの近傍データへの距離の 3 種類の距離を重みで加算平均した構成とする．

このように，pdi-Bagging では，バーチャルデータを発生してそのクラスを推定し，学習データに追加して，ファジィ推論 [30, 31] の弱判別器で判別線を推定し，次層でも同様にバーチャルデータの追加とクラス推定，及び，学習データへの追加で判別線の推定を行う．これらの一連の操作を繰り返し，最終的には，複数の弱判別器の多数決によって評価データの識別率を得る．ファジィ理論を用いたアンサンブル法としては，ファジィ意思決定によるアンサンブル法 Fuzzy Adaptive Boosting [32] が提案されており，顔認識画像の事例により，Fuzzy Adaptive Boosting が AdaBoost よりも識別率が高いことを紹介している．また，ニューロファジィシステムを用いたアンサンブル法 [33] も提案されており，AdaBoost の識別率よりも高い識別精度を示している．さらに，ファジィ推論を用いたアンサンブル法 [34] やファジィランダムフォレスト [35] も提案されている．しかし，これらのアンサンブル法では，学習法の定式化のみが議論されており，学習データの均一性やデータの偏りを解消する手法は議論されていない．pdi-Bagging は，バーチャルデータを発生し学習データに追加するので，データの偏りが解消され，学習データの均一性が保たれる特性がある．ここでは，新たな pdi-Bagging のアルゴリズムを定式化し，数値例を用いて本手法の有用性を議論する．

2. pdi-Bagging

バギングとは，複数個の弱判別器を用意し，各識別結果を統合することにより評価データに対する高い識別率を得る手法である．ブースティングが繰り返し学習の際に，複数の学習データ間で相互依存関係を有するのに対して，バギングでは，複数の学習データ間は独立である．pdi-Bagging では，この弱判別器の独立性に加えて，バーチャルデータの発生により学習データの量を増加させる．

pdi-Bagging の概念図を図 1 に示す．pdi-Bagging では，まず，全データ集合から確率的に抽出された学習データ (TRD) を用いてファジィ推論の弱判別器 M_0 を学習し TRD の識別率を算出する．次ステップ (層) では，メンバシップ関数を用いて，特定の学習データからバーチャルデータを発生し TRD の量を増加して，ファジィ推論の弱判別器 M_1 により TRD の識別率を算出する．TRD の量を増加させることにより弱判別器の識別精度が向上する．終了判定が満足されるまでこの一連の操作を L 回繰り返し，最終的に，評価データ (CHD) を L 個の弱判別器 $M_0, M_1, \cdots, M_l, \cdots, M_L$ に入力して，多数決により最終結果を得る．pdi-Bagging では，バーチャルデータを学習データに追加し弱判別器が同定されるので，従来のバギングや AdaBoost よりも高い精度の判別線が得られる [26].

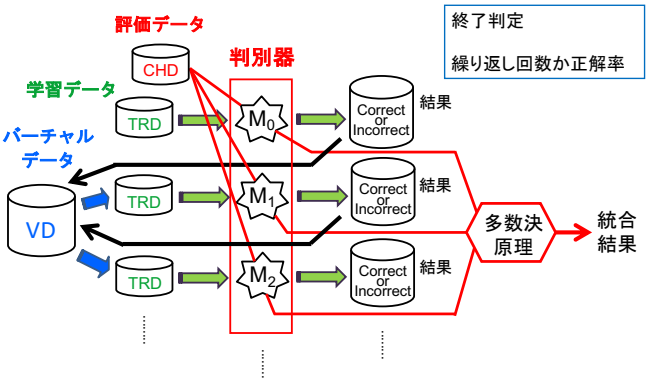


図 1　pdi-Bagging Algorithm

pdi-Bagging では，弱判別器として簡易型ファジィ推論 [31] によるファジィクラスタリングを用いる．特に，弱判別器をファジィクラスタリングに限定する必要はない．しかし，ファジィ推論は学習能力と表現性に優れ，ルール表現によって，深層学習等の課題である結果の可視化を実現できる．また，ファ

ジィ推論は決定木の一手法であり，ファジィ集合によって判別線をより柔軟に学習することも可能である．これらの利点から，ここでは，ファジィ推論を用いる．簡易型ファジィ推論は，if-then 型でルールを表現し，前件部では，メンバシップ関数のファジィ集合を定義し，後件部では，シングルトンを定義する．ここでは，三角型のメンバシップ関数を一般化した正規な台形型ファジィ集合を用いる．

いま，出力変数を z，後件部のシングルトンを p_i で表すと，ファジィルール r_i, $i = 1, 2, \cdots, R$ は次のようになる．

$$r_i : \text{if} \;\; x_1 \text{ is } \mu_{F_{i1}}(x_1) \text{ and } \;\cdots\; \text{and } x_n \text{ is } \mu_{F_{in}}(x_n)$$
$$\text{then} \;\; C = \{C_{ik} \mid z = p_i\}$$

ただし，C は出力クラスの変数であり，C_{ik} はルール r_i のクラス値が C_k であることを示す．

いま，入力データ $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ が得られたとしよう．第 i 番目のファジィルール r_i の前件部に入力データ $\mathbf{x}$ を入力し，前件部の適合度 $\mu_i(\mathbf{x}) = \mu_{F_{i1}}(x_1) \cdot \mu_{F_{i2}}(x_2) \cdot \;\cdots\; \cdot \mu_{F_{in}}(x_n)$ を計算する．ファジィ推論の結果 $\hat{z}$ とクラス C は次式から求める．

$$\hat{z} = \frac{\sum_{i=1}^{R} \mu_i(\mathbf{x}) \cdot p_i}{\sum_{i=1}^{R} \mu_i(\mathbf{x})}$$
$$C = \{C_k \mid \min |\hat{z} - z|\}$$

次に，pdi-Bagging におけるバーチャルデータの生成方法について説明する．いま，W 個のデータからなるデータ集合 D の第 d 番目のデータを $\mathbf{x}^D(d) = (x_1^D(d), x_2^D(d), \cdots, x_j^D(d), \cdots, x_n^D(d))$ で表す．バーチャルデータ $\mathbf{x}^V(d)$ は，ある特定の正判別データ $\mathbf{x}^C(d)$ や誤判別データ $\mathbf{x}^E(d)$ の周辺に発生する．$\mathbf{x}^V(d)$ を構成する第 d 番目のデータの第 j 属性目のバーチャルデータ $x_j^V(d)$ の発生は，ある実数 h, $0 \le h \le 1$ が与えられると，ファジィ数 F のメンバシップ関数 $\mu_F(x_j)$ を用いて次のように発生する．

$$x_j^V(d) = \{x_j \mid \mu_F(x_j) = h,\ \mu_F(x_j^S(d)) = 1\}$$
$$h \sim N(1, 1), \qquad 0 \le h \le 1$$

ただし，$x_j^S(d)$ は正判別データ $x_j^C(d)$ または誤判別データ $x_j^E(d)$ を意味し，メンバシップ関数 $\mu_F(x_j)$ は，ファジィ数 F の中心が $x_j^S(d)$ であり，標準偏差が σ である次の正規分布で定義する．

$$\mu_F(x_j) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_j - x_j^S(d))^2}{2\sigma^2}\right)$$

バーチャルデータの発生方法として次の 5 種類を提案する．

(1) 正判別全領域型 (CA: Generation of Virtual Data with Correct Data in All Area)

全領域の任意の学習データ $\mathbf{x}^S(d)$ のクラスが弱判別器により正判別と判別された場合，その正判別データ $\mathbf{x}^C(d)$ の周辺にバーチャルデータ $\mathbf{x}^V(d)$ を発生する．

(2) 正判別クラスター中心型 (CC: Generation of Virtual Data with Correct Data arround Cluster Center)

任意の学習データ $\mathbf{x}^S(d)$ のクラスが弱判別器により誤判別と判別された場合，その誤判別データ $\mathbf{x}^E(d)$ の真のクラス (教師クラス) と同クラスに属する $\mathbf{x}^E(d)$ からの最近傍の正判別データと最遠方の正判別データの中点を求め，その中点に最近傍の正判別データ $\mathbf{x}^C(d')$ の周辺にバーチャルデータ $\mathbf{x}^V(d')$ を発生する．

$$\mathbf{x}^C(d') = \{\mathbf{x}^C(e) \mid \min_e |\mathbf{x}^C(e) - \frac{1}{2}(\max_f |\mathbf{x}^E(d) - \mathbf{x}^C(f)|$$
$$+ \min_g |\mathbf{x}^E(d) - \mathbf{x}^C(g)|)|, \quad for \;\; \forall e, f, g\}$$

(3) 誤判別型 (E: Generation of Virtual Data with Error Data)

任意の学習データ $\mathbf{x}^S(d)$ のクラスが弱判別器により誤判別と判別された場合，その誤判別データ $\mathbf{x}^E(d)$ の周辺にバーチャルデータ $\mathbf{x}^V(d)$ を発生する．

(4) 混合判別全領域型 (MA: Generation of Virtual Data with Correct/Error(Mix) Data in All Area)

バギングの各層で正判別全領域型と誤判別型を交互に用いて，バーチャルデータ $\mathbf{x}^V(d)$ を $\mathbf{x}^C(d)$ や $\mathbf{x}^E(d)$ の周辺に発生する．

(5) 混合判別クラスター中心型 (MC: Generation of Virtual Data with Correct/Error(Mix) Data arround Cluster Center)

バギングの各層で正判別クラスター中心型と誤判別型を交互に用いて，バーチャルデータ $\mathbf{x}^V(d)$ を $\mathbf{x}^C(d)$ や $\mathbf{x}^E(d)$ の周辺に発生する．

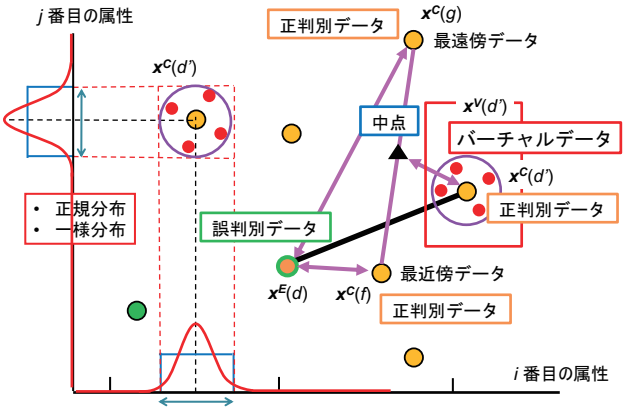


図 2　Generation of Virtual Data with Correct Data arround Cluster Center

特に，正判別クラスター中心型 (CC) について，図 2 を用いて概説する．図 2 では，合計 8 個のデータを緑色と黄色の 2 クラスに識別する問題を想定している．図中下部の黄色に緑色の枠を有する学習データは，教師クラスは黄色クラスであるが，緑色クラスとして誤判別されている．この誤判別データ $\mathbf{x}^E(d)$ は，教師クラスが黄色クラスであるので，最近傍データ $\mathbf{x}^C(f)$ と最遠方データ $\mathbf{x}^C(g)$ を求め，それらの中点 (図中では ▲ 印)

を計算し，この中点の最近傍の正判別データ $\mathbf{x}^C(d')$ を求めて，その周辺にバーチャルデータ $\mathbf{x}^V(d')$ を発生する．そのため，バーチャルデータはクラスター中心付近で発生する傾向があり，正判別クラスター中心型 (CC) は正判別全領域型 (CA) と比較して，バーチャルデータが判別線に影響を与えやすい．また，中点座標を最近傍と最遠方の任意の内挿点や外挿点に変更することで，判別線に与える影響を制御することも可能となる．

3. クラス変更化と定式化

pdi-Bagging では，誤判別データや正判別データの周辺にバーチャルデータを発生して識別率を向上させる．しかし，さらに識別率を向上させるため，発生したバーチャルデータのクラスを正しく評価する必要がある．ここでは，バーチャルデータに正しいクラスを付与するためのクラス決定法を提案する．

いま，正判別データ $\mathbf{x}^C(d)$ や誤判別データ $\mathbf{x}^E(d)$ を発生源としてバーチャルデータ $\mathbf{x}^V(d)$ を発生させたとする．基本的には，$\mathbf{x}^V(d)$ のクラスは発生源の $\mathbf{x}^S(d) = \{\mathbf{x}^{C,k}(d),\ \mathbf{x}^{E,k}(d)\}$ の教師クラスと同クラスとすべきと考える．しかし，バーチャルデータは発生源データから離れた位置に発生する場合もあり，また，異クラスの学習データ領域内でバーチャルデータが発生することもある．そこで，バーチャルデータ $\mathbf{x}^V(d)$ のクラス k^* を次の 3 つの評価基準：正誤判別データの評価 (E_1)，識別クラスの評価 (E_2)，近傍データクラスの評価 (E_3) を要素として統合の評価式により決定する．

(1) 正誤判別データの評価 (E_1)

評価値 E_1 は，バーチャルデータ $\mathbf{x}^V(d)$ とクラス k をもつ発生源の正誤判別データ $\mathbf{x}^{S,k}(d)$ との距離を用いて定義する．この評価値 E_1 が小さいほど，$\mathbf{x}^V(d)$ はクラス k への依存度が高い．

$$E_1^k = \frac{|\mathbf{x}^V(d) - \mathbf{x}^{S,k}(d)|}{\max_e |\mathbf{x}^{S,k}(d) - \mathbf{x}^{D+V}(e)| - \min_f |\mathbf{x}^{S,k}(d) - \mathbf{x}^{D+V}(f)|}, \quad for\ \ \forall e, f$$

$$E_1^p = 1 - E_1^k, \quad for\ \ p \neq k$$

(2) 識別クラスの評価 (E_2)

評価値 E_2 は，バーチャルデータ $\mathbf{x}^V(d)$ とクラス k の中心との距離を用いて定義する．この評価値 E_2 が小さいほど，$\mathbf{x}^V(d)$ はクラス k への依存度が高い．いま，クラス k の中心を $\mathbf{x}_c^k$ とする．

$$E_2^k = \frac{|\mathbf{x}^V(d) - \mathbf{x}_c^k|}{\max_{e,f} |\mathbf{x}^{D+V}(e) - \mathbf{x}^{D+V}(f)|}, \quad for\ \ \forall e, f$$

(3) 近傍データクラスの評価 (E_3)

評価値 E_3 は，バーチャルデータ $\mathbf{x}^V(d)$ とクラス k の最近傍の正誤判別データ $\mathbf{x}^{S,k}(e)$ との距離を用いて定義する．この評価値 E_3 が小さいほど，$\mathbf{x}^V(d)$ はクラス k への依存度が高い．

$$E_3^k = \frac{\min_e |\mathbf{x}^V(d) - \mathbf{x}^{S,k}(e)|}{\max_{f,g} |\mathbf{x}^{D+V}(f) - \mathbf{x}^{D+V}(g)|}, \quad for\ \ \forall e, f, g$$

これらの評価基準では，バーチャルデータが発生源データの近傍で発生する場合には評価 E_1 が高まり，クラスの中心近傍で発生する場合には評価 E_2 が高まる．また，近傍データのクラスへの評価は E_3 で計算される．

これらの 3 つの評価基準を統合し全体の評価値 E^k を得る．全体の評価値は，$\mathbf{x}^V(d)$ のクラスとして下記の評価値 E^k が最小となるクラス k^* を求める．

$$k^* = \{k | \min_k E^k = \min_k (w_1 E_1^k + w_2 E_2^k + w_3 E_3^k)\} \qquad (1)$$

ただし，w_1, w_2, w_3 は各評価値の重みである．

pdi-Bagging のアルゴリズムを次のように定式化する．

Step 1 計測データ D(個数：W 個) を学習データ D^{TRD}(個数：W^{TRD} 個) と評価データ D^{CHD}(個数：W^{CHD} 個) に分割する．また，D^{TRD} から構成されるバーチャルデータを D^V で表す．

Step 2 第 l 番目の弱判別器 M_l に D^{TRD} を入力し，第 l 番目の結果 R_l の識別率 r_l^{TRD} を得る．ただし，M_0 は初期弱判別器である．

Step 3 正判別あるいは誤判別された第 d 番目のデータを D^{TRD} から一時的に抽出する．正判別あるいは誤判別データ $\mathbf{x}^S(d)$ の第 j 番目の属性値 $x_j^S(d)$ に対して，メンバシップ関数 $\mu_F(x_j)$ によりバーチャルデータ $x_j^V(d)$ を発生させる．

Step 4 式 (1) により，バーチャルデータ $\mathbf{x}^V(d)$ のクラス k^* を求める．$l > 2$ の第 $l-1$ 番目の D^V からバーチャルデータ $\mathbf{x}^V(d)$ を除去し，第 l 番目の D^V にクラス k^* をもつバーチャルデータ $\mathbf{x}^V(d)$ を追加する．

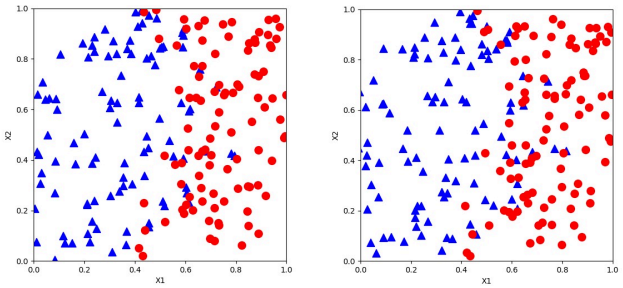
Step 5 乱数により D^V から v 個のバーチャルデータを取り出し D^{TRD} に加える．

Step 6 $l = l + 1$ として Step2 から Step5 までを繰り返し，しきい値 θ に対して $r_l^{TRD} \geq \theta$ を満足した $K = l$ の時点，あるいは，弱判別器の個数 L と繰り返し回数 K，$K \leq L$ に対して $l \geq K$ を満足した時点でアルゴリズムを終了する．

Step 7 M_0，M_1，$\cdots$，M_l，$\cdots$，M_K に D^{CHD} を適用し，多数決により結果の識別率 r_K^{CHD} を得る．

4. 数値データによる検証と考察

検証に用いる数値データは，学習データと評価データでそれぞれ 200 個である．学習データと評価データに用いる数値データを図 3 に示す．これらの数値データは，基本データに乱数で最大 ±0.05 の増減値を加えて生成した．ここでは，2 入力 1 出力の 2 群判別問題として，ファジィ推論の 2 群クラスの実数値を 2.0（赤・○印）と 3.0（青・△印）に設定した．なお，ここでは，2 fold cross validation を用いるので，数値例 1 を学習データ，数値例 2 を評価データとして識別率を求め，逆のパターンでも識別率を求め，それらを平均識別率として求めた．



(a) 数値例 1　　(b) 数値例 2

図 3　Numerical Example of Training and Testing Data

弱判別器には簡易型ファジィ推論を用い，まず各入力区間 [0, 1] に 5 種類の台形型メンバシップ関数を設定し，データ空間の全領域に 25 個のルールを設定した．次に，データ空間の特定領域にルールを追加した場合の識別率の変化を確認するため，特定領域を G_1，G_2 として，$G_1 = \{(x_1, x_2) \mid [0.4, 0.7] \times [0.4, 0.7]\}$ に 49 個のルールを追加し，また，$G_2 = \{(x_1, x_2) \mid [0.7, 0.8] \times [0.3, 0.7]\}$ に 4 個のルールを追加した．その結果，ルール数は合計 78 個となった．この特定領域へのルールの追加は，判別線が複雑な境界面であるとき，判別線から離れた特異点データの集合領域に別途，追加ルールを構成した際の精度を検証するためである．ここでは，追加ルールの有無と特定領域内の各次元の両端でメンバシップ関数を台形型と直角台形型で定義した場合の合計 3 種類を設定した．特定領域内のメンバシップ関数は台形型とするが，特定領域内の各次元の両端に直角台形型メンバシップ関数を設定した場合には，特定領域はメンバシップ関数の学習後であっても領域は変化しない．一方，特定領域内の各次元の両端に台形型メンバシップ関数を設定した場合には，学習によって領域は変化する．したがって，追加ルールに直角台形型メンバシップ関数を設定した場合には，メンバシップ関数は学習によって特定領域外に移動されず，特定領域内で集中的に学習される．

ファジィ推論の前件部の初期値設定は既定法とし，前件部と後件部の学習順序は，後件部→前後件部交互学習とした．学習では，各入力の台形型メンバシップ関数の上底の 2 頂点の x 座標 x_b と x_c，および，上底と下底の x 座標の差 α と β の学習係数 K_b，K_α，K_c，K_β [30] を同一とし 0.01 に設定した．また，後件部のシングルトンの学習係数 K_p は，最初の後件部学習では 0.4 に設定し，交互学習の後件部学習では 0.6 とした．後件部と前後件部交互学習のエポック回数をそれぞれ 10 回と (10, 10) 回に設定した．

バーチャルデータ発生のメンバシップ関数 $\mu_F(x_j)$ は正規分布とし，バーチャルデータの発生個数は 1 個とした．ただし，事前実験でファジィ推論の識別率が 87% 程度と得られ，評価データの 200 個で 26 個程度が誤判別となることから，バーチャルデータの総数が学習データの 200 個と同程度となるためには，バーチャルデータの発生個数は 8 個程度となる．そこで，バーチャルデータの発生個数を 1 個から 10 個まで変化した場合の識別率もあわせて議論した．

バーチャルデータのクラス推定のための評価値の重みを $(w_1, w_2, w_3) = \{(1/3, 1/3, 1/3), (0.2, 0.4, 0.4), (0.2, 0.3, 0.5), (0.2, 0.5, 0.3), (0.5, 0.25, 0.25), (0.01, 0.495, 0.495), (0.05, 0.475, 0.475)\}$ とする．重みの決定では，発生源データからの距離の重み w_1 がクラス推定に最も大きな影響を与えるので，$w_1 = w_2 = w_3$ の $w_1 = 1/3$ の場合と $w_1 = 0.5$ の場合，および，w_1 の値を減じた 5 種類の合計 7 種類で識別率を議論した．

アルゴリズムの終了規範は繰り返し判定として，回数は $K = 5$ とした．ただし，混合判別型は，奇数層を誤判別型とし，偶数層を正判別型とする．ファジィ推論の後件部と前後件部交互学習では，1 エポックごとに乱数によりデータの並び順序を変更し，エポック回数は，それぞれ 10 回と (10, 10) 回であるので，$K = 5$ では合計 150 回のエポック学習となる．ここでは，2-fold cross validation を用いるので，それぞれのデータ集合の 150 回のエポック学習は合計 300 回のエポック学習となり，これを 1 試行として，正判別全領域型，正判別クラスター中心型，誤判別型，混合判別全領域型，混合判別クラスター中心型で，それぞれ 10 試行で得た識別率の平均値を比較した．

正判別全領域型 (CA)，正判別クラスター中心型 (CC)，誤判別型 (E)，混合判別全領域型 (MA)，混合判別クラスター中心型 (MC) の評価データに対する識別率を表 1 と図 4〜図 6 に示す．表 1 では，メンバシップ関数とルール数の種別によるバーチャルデータ発生手法の識別率を示し，同時に，メンバシップ関数とルールの種別の差も計算した．図 4〜図 6 では，メンバシップ関数とルール数の種別による平均識別率を示した．

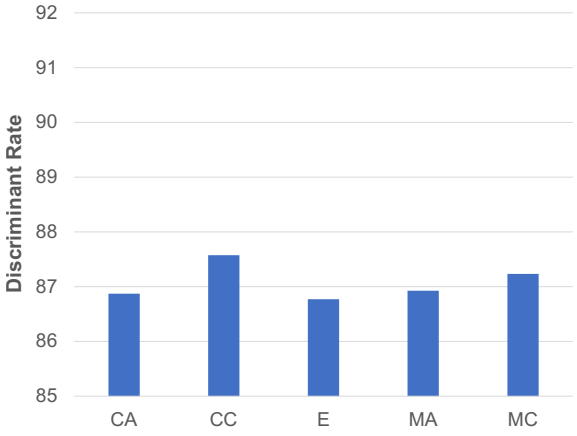


図 4　Average Discriminant Rates of 5 Methods in 25 Basic Rules

まず，表 1 と図 4 の結果から，台形型メンバシップ関数による 25 個のルール数の場合，識別率には次の特性がある．

1) 25 ルール数の単体のファジィ推論の 2-fold cross validation による識別率は 84.40% となった．5 種類のすべてのバーチャルデータ発生法の識別率は単体のファジィ推論より上回っており，本手法の有効性が示されている．
2) 正判別型の 2 手法の比較では，全領域型 (CA) は平均識別率が 86.87% であるのに対して．クラスター中心型 (CC) は全領域型 (CA) よりも全般的に識別率が高く，平均識別

表 1 Comparison of Discriminant Rates According to 5 Methods

ルール形式	評価指標の重み	正・全 (CA) (%)			正・ク (CC) (%)			誤 (E) (%)			混・全 (MA) (%)			混・ク (MC) (%)		
		識別率	差 (a)	差 (b)	識別率	差 (a)	差 (b)	識別率	差 (a)	差 (b)	識別率	差 (a)	差 (b)	識別率	差 (a)	差 (b)
(a) 台形型 MF 25 ルール	1/3, 1/3, 1/3	86.73	—	—	87.70	—	—	86.61	—	—	87.05	—	—	87.29	—	—
	0.2, 0.4, 0.4	86.50	—	—	87.60	—	—	87.03	—	—	87.28	—	—	87.52	—	—
	0.2, 0.3, 0.5	87.00	—	—	87.55	—	—	86.70	—	—	87.03	—	—	87.10	—	—
	0.2, 0.5, 0.3	86.85	—	—	87.70	—	—	86.70	—	—	87.08	—	—	87.15	—	—
	0.5, 0.25, 0.25	86.40	—	—	87.45	—	—	86.95	—	—	86.85	—	—	87.40	—	—
	0.01, 0.495, 0.495	87.18	—	—	87.55	—	—	86.55	—	—	86.58	—	—	86.88	—	—
	0.05, 0.475, 0.475	87.45	—	—	87.48	—	—	86.85	—	—	86.63	—	—	87.30	—	—
	平均	86.87	—	—	87.58	—	—	86.77	—	—	86.93	—	—	87.23	—	—
(b) 台形型 MF 78 ルール	1/3, 1/3, 1/3	89.53	2.80	—	89.83	2.13	—	89.80	3.18	—	89.78	2.73	—	89.79	2.50	—
	0.2, 0.4, 0.4	89.33	2.83	—	89.93	2.33	—	90.15	3.13	—	90.00	2.73	—	89.95	2.43	—
	0.2, 0.3, 0.5	89.03	2.03	—	90.15	2.60	—	90.30	3.60	—	89.78	2.75	—	89.93	2.83	—
	0.2, 0.5, 0.3	88.95	2.10	—	89.65	1.95	—	90.05	3.35	—	89.85	2.78	—	89.83	2.67	—
	0.5, 0.25, 0.25	89.18	2.77	—	89.48	2.03	—	90.05	3.10	—	89.38	2.53	—	90.23	2.83	—
	0.01, 0.495, 0.495	87.40	0.22	—	89.80	2.25	—	88.63	2.08	—	88.55	1.97	—	89.43	2.55	—
	0.05, 0.475, 0.475	88.70	1.25	—	90.00	2.52	—	89.83	2.97	—	89.85	3.23	—	89.85	2.55	—
	平均	88.87	2.00	—	89.83	2.26	—	89.83	3.06	—	89.60	2.67	—	89.86	2.62	—
(c) 直角台形型 MF 78 ルール	1/3, 1/3, 1/3	90.03	3.30	0.50	90.33	2.63	0.50	89.93	3.32	0.14	90.23	3.18	0.45	90.15	2.86	0.36
	0.2, 0.4, 0.4	89.83	3.33	0.50	90.20	2.60	0.27	90.35	3.33	0.20	90.28	3.00	0.27	90.28	2.76	0.32
	0.2, 0.3, 0.5	90.45	3.45	1.43	90.10	2.55	-0.05	90.05	3.35	-0.25	90.10	3.08	0.32	90.30	3.20	0.37
	0.2, 0.5, 0.3	89.95	3.10	1.00	90.35	2.65	0.70	90.05	3.35	0.00	90.30	3.23	0.45	89.98	2.82	0.15
	0.5, 0.25, 0.25	90.18	3.78	1.00	90.28	2.83	0.80	89.93	2.97	-0.13	90.05	3.20	0.67	90.35	2.95	0.13
	0.01, 0.495, 0.495	87.55	0.37	0.15	90.40	2.85	0.60	88.63	2.07	0.00	88.40	1.82	-0.15	90.18	3.30	0.75
	0.05, 0.475, 0.475	89.83	2.37	1.13	90.35	2.87	0.35	89.95	3.10	0.12	90.03	3.40	0.17	90.03	2.73	0.17
	平均	89.69	2.81	0.81	90.29	2.71	0.45	89.84	3.07	0.01	89.91	2.99	0.31	90.18	2.94	0.32

率は 87.58% であった.

3) 誤判別型の識別率は正判別型よりも必ずしも高いとはいえない. 平均識別率は 86.77% であった.
4) 混合判別型の 2 手法の比較では, 識別率は混合判別全領域型 (MA) も混合判別クラスター中心型 (MC) もほぼ同じであるが, 平均識別率は混合判別クラスター中心型 (MC) の方が混合判別全領域型 (MA) よりも若干高い.

台形型メンバシップ関数によるルール数が 25 個の場合には, 5 手法の比較では, 正判別クラスター中心型 (CC) の識別率は正判別全領域型 (CA) よりも高く, 混合判別型でも, 混合判別クラスター中心型 (MC) の方が, 混合判別全領域型 (MA) よりも識別率は高い. この理由は, バーチャルデータがクラスターの中心付近で発生するので, クラス中心付近のファジィルールが精度良く学習されることによる.

次に, 表 1 と図 5 の結果から, 特定領域内で台形型メンバシップ関数による追加ルールを加えた合計 78 個の場合について, 識別率には次の特性があることがわかる.

1) 78 ルール数の台形型メンバシップ関数による単体のファジィ推論の 2-fold cross validation の識別率は 89.68% となった. 5 種類のバーチャルデータ発生法の中で, 誤判別型 (E) と正判別クラスター中心型 (CC), 混合判別クラスター中心型 (MC) は, 単体のファジィ推論の識別率より上回っており, 全領域でバーチャルデータを発生しない手法の有効性が示されている.
2) 正判別型の 2 手法の比較では, 識別率はクラスター中心型 (CC) が全領域型 (CA) よりも高い. 特に, 全領域型 (CA) の平均識別率は 88.87% であったが, クラスター中心型 (CC) の平均識別率は 89.83% であった.
3) 誤判別型 (E) の識別率は正判別型と同等である. しかし, 識別率の最大値は正判別型よりも高い.
4) 混合判別型の 2 手法の比較では, 混合判別全領域型 (MA) と混合判別クラスター中心型 (MC) の識別率はほぼ同じであるが, 混合判別クラスター中心型 (MC) の平均識別率はわずかに高く 89.86% であった.
5) 識別率の増減差から, 台形型メンバシップ関数の 25 個ルールとの比較では, 全ての手法において, 平均識別率は 2.00%〜3.06% ほど上昇した. ただし, 正判別全領域型 (CA) と正判別クラスター中心型 (CC) の平均識別率の上昇率は他手法よりも若干低い.

図 5 Average Discriminant Rates of 5 Methods in 78 Total Rules Added by Trapezoidal Membership Function

表 2　Results of t-Test and Tukey's HSD Test between 5 Methods in 25 Basic Rules

検定方法 バーチャルデータ 発生方法	識別率	t 検定 正・全 (CA)	 正・ク (CC)	 誤 (E)	 混・全 (MA)	 混・ク (MC)	Tukey's HSD 検定 正・全 (CA)	 正・ク (CC)	 誤 (E)	 混・全 (MA)	 混・ク (MC)	識別率	Tukey's HSD 検定・2 fold cross validation 正・全 (CA)	 正・ク (CC)	 誤 (E)	 混・全 (MA)	 混・ク (MC)
正・全 (CA)	87.19 86.55	— —	①★ 0.3557	①★ ②	①★ ②	①★ ②	— —	①★ 0.9870	①★ ②	①★ 0.1384	①★ 0.1042	86.87	—	◯★	0.9282	0.9922	0.0605
正・ク (CC)	88.70 86.45	① 0.3557	— —	① ②	① ②	0.0514 ②	① 0.9870	— —	① ②	① 0.3293	0.4949 0.2627	87.58	◯	—	◯	◯	0.0846
誤 (E)	87.85 85.69	① ②★	①★ ②★	— —	0.3725 ②★	①★ ②★	① ②★	①★ ②★	— —	0.9949 0.3575	①★ 0.4361	86.77	0.9282	◯★	—	0.7313	◯★
混・全 (MA)	87.79 86.06	① ②★	①★ ②★	0.3725 ②	— —	①★ 0.4165	① 0.1384	①★ 0.3293	0.9949 0.3575	— —	①★ 0.9999	86.93	0.9922	◯★	0.7313	—	0.1483
混・ク (MC)	88.43 86.04	① ②★	0.0514 ②★	① ②	① 0.4165	— —	① 0.1042	0.4949 0.2627	① 0.4361	① 0.9999	— —	87.23	0.0605	0.0846	◯	0.1483	—

台形型メンバシップ関数によるルール数が 78 個の場合，正判別型，誤判別型，混合判別型の比較では，識別率はほぼ同じであるが，平均識別率は，混合判別クラスター中心型 (MC) が最も良い結果となった．一方，ルール数が 25 個の場合と比較すると，全般的に識別率が高くなり，平均識別率が 2.52% ほど上昇した．ルール数が追加されたので識別率の上昇は予測できるが，ルールを追加した特定領域は特異点データの集合領域周辺であるので，その特定領域周辺の識別率の向上が全体の識別率を押し上げている効果が現れていると考えられる．

図 6　Average Discriminant Rates of 5 Methods in 78 Total Rules Added by Right Trapezoidal Membership Function

さらに，表 1 と図 6 の結果から，特定領域内の両端で直角台形型メンバシップ関数による追加ルールを加えた合計 78 個の場合について，識別率には次の特性があることがわかる．

1) 78 ルール数の直角台形型メンバシップ関数による単体のファジィ推論の 2-fold cross validation の識別率は 89.73% となった．5 種類のバーチャルデータ発生法の中で，誤判別型 (E) と混合判別全領域型 (MA)，混合判別クラスター中心型 (MC)，正判別クラスター中心型 (CC) は，単体のファジィ推論の識別率より上回っており，特に，混合判別クラスター中心型 (MC) と正判別クラスター中心型 (CC) は 0.45% 以上高い．
2) 正判別型の 2 手法の比較では，識別率は，クラスター中心型 (CC) が全領域型 (CA) よりもわずかながら高い．
3) 誤判別型 (E) の識別率は正判別全領域型 (CA) と同等である．
4) 混合判別型の 2 手法の比較では，混合判別クラスター中心型 (MC) が混合判別全領域型 (MA) よりも平均識別率がわずかに高い 90.18% であった．
5) 識別率の増減差から，台形型メンバシップ関数の 25 個ルールとの比較では，全ての手法において，平均識別率は 2.71%〜3.07% ほど上昇した．ただし，正判別全領域型 (CA) と正判別クラスター中心型 (CC) の平均識別率の上昇率は他手法よりも若干低い．

直角台形型メンバシップ関数によるルール数が 78 個の場合，正判別型，誤判別型，混合判別型の比較では，識別率はほぼ同じであるが，平均識別率は，正判別クラスター中心型 (CC) と混別判別クラスター中心型 (MC) が良い結果となった．一方，ルール数が 25 個の場合と比較すると，全般的に識別率が高くなり，平均識別率が 2.90% ほど上昇した．また，台形型でルール数が 78 個の場合と比較すると，わずかながら高くなり，平均識別率が 0.38% ほど上昇した．平均識別率が 90.00% 以上であるのは，正判別クラスター中心型 (CC) の 90.29% と混合判別クラスター中心型 (MC) の 90.18% であった．ルールを追加した特定領域は特異点データが多く存在する領域に設定しているので，この領域での識別率の向上が全体の識別率を押し上げている効果が現れていると考えられ，さらに，特定領域のメンバシップ関数として直角台形を定義した場合には，メンバシップ関数の学習でも特定領域のサイズは変化しないので，特定領域内で効率的にメンバシップ関数が学習され，識別率が上昇していると考えられる．

台形型メンバシップ関数のルール数が 25 個での 5 種類の手法を比較するため，評価指標の重みの平均識別率を用いた t 検定と Tukey's HSD 検定の結果を表 2 に示す．表 2 の左列の t 検定と中央列の Tukey's HSD 検定では，図 3 の数値データを交互に学習データと評価データに用いて，5 種類の手法の識別率の組み合わせが有意水準 5% の片側 t 検定と Tukey's HSD 検定で有意な差を認めた場合，データ集合の交互の組に対して ① と ② で示した．さらに，手法間で有意な差があり識別率が高い手法には，その手法の列に ★ を付与した．なお，① と ② の

検定結果 p が有意水準 5% 以下を満たさない場合には，その p 値を示した．したがって，① または ② が示され ★ が付与されている列の手法は，他手法とは有意な差が認められる識別率の高い手法といえる．一方，表 2 の右列の Tukey's HSD 検定・2-fold cross validation では，図 3 の数値データを 2-fold cross validation として，5 種類の手法の識別率の組み合わせが有意水準 5% の Tukey's HSD 検定で有意な場合，その組み合わせに対して ◯ で示した．さらに，手法間で有意な差があり識別率が高い手法には，その手法の列に対して ★ を付与した．なお，同様に，検定結果 p が有意水準 5% 以下を満たさない場合，その p 値を示した．したがって，◯ が示され ★ が付与されている列の手法は，他手法とは有意な差が認められる識別率の高い手法といえる．最後に，左列の t 検定と中央列の Tukey's HSD 検定の各手法の列において，★ 印が付与された ① または ② の最大個数の列と第 2 番目の列を灰色で示し，右列の Tukey's HSD 検定・2-fold cross validation において，★ が付与された ◯ 印の最大個数の列と第 2 番目の列を灰色で示した．

表 3　Comparison of Discriminant Rates between 25 Rules and 78 Rules

ルール形式	正・全 (CA)	正・ク (CC)	誤 (E)	混・全 (MA)	混・ク (MC)
台形型 MF・25 ルール	86.87143	87.57500	86.76981	86.92500	87.23395
台形型 MF・78 ルール	88.87143	89.83214	89.82792	89.59643	89.85625
直角台形型 MF・78 ルール	89.68571	90.28571	89.84026	89.91104	90.17857

表 4　Two-way Analysis of Variance Table between Rule Formats and Virtual Data Generation Methods

変動要因	変動	自由度	分散	観測分散比	P 値	F 境界値
MF・ルール形式	24.91134	2	12.45567	276.62	4.1282E-08	4.45897
VD 発生方法	1.01819	4	0.25455	5.6531	0.018445	3.83785
誤差	0.36022	8	0.04503			
合計	26.28975	14				

これらの t 検定と Tukey's HSD 検定の結果から，灰色の列の手法は，すべて正判別クラスター中心型 (CC) と混合判別クラスター中心型 (MC) であり，したがって，これらの手法は，他手法よりも有意差が認められる識別率の高い手法であるといえる．特に，右列の Tukey's HSD 検定・2-fold cross validation では，◯ 印と ★ 印が同時に示されている手法は，灰色の正判別クラスター中心型 (CC) と混合判別クラスター中心型 (MC) のみであり，◯ 印以外の p 値も 0.0846 と 0.0605, 0.1483 となっており，他の手法間の p 値よりもかなり小さい．また，表 1 から，正判別全領域型 (CA)，誤判別型 (E)，混合判別全領域型 (MA) は識別率が低く，正判別クラスター中心型 (CC) と混合判別クラスター中心型 (MC) は識別率が高い．したがって，これらの結果から，正判別クラスター中心型 (CC) と混合判別クラスター中心型 (MC) は他の手法よりも識別率が高い手法であるといえる．

一方，台形型メンバシップ関数の 25 個ルールと 78 個ルールの識別率を表 3 に示す．また，メンバシップ関数の形状・ルール数とバーチャルデータ発生手法を比較するため，二元配置分散分析の結果を表 4 に示す．表 4 は表 3 の識別率を二元配置分散分析法で検定した結果である．さらに，表 5 に t 検定と Tukey's HSD 検定の結果を示す．いま，表 4 において，帰無仮説として，表 3 の因子 (行) の各メンバシップ関数の形状とルール数の識別率が等しく，因子 (列) の各バーチャルデータ発生手法の識別率が等しいと仮定する．表 4 の有意水準を 5% とした分散分析の結果では，第 1 行目の p 値は 4.1282E-08 と得られた．この結果から，表 3 の因子 (行) の各メンバシップ関数の形状とルール数の識別率には，5% の有意水準で有意な差があるといえる．また，第 2 行目の p 値は 0.018445 と得られた．したがって，表 3 の因子 (列) のそれぞれのバーチャルデータ発生手法の識別率には，5% の有意水準で有意な差があるといえる．一方，表 5 における台形型メンバシップ関数の 25 個ルールと 78 個ルールの比較では，有意水準 5% の片側 t 検定と Tukey's HSD 検定で全ての手法間で有意な差が認められた．また，台形型メンバシップ関数の 25 個ルールと直角台形型メンバシップ関数の 78 個ルールの比較でも，有意水準 5% の片側 t 検定と Tukey's HSD 検定で全ての手法間で有意な差が認められた．これらの結果から，台形型と直角台形型のメンバシップ関数の形状の違いや 25 ルールと 78 ルールのルール数の違い，および，バーチャルデータ発生手法の違いによって，検定の結果からも識別率に明らかな差があると認められた．したがって，メンバシップ関数を直角台形型として，特定領域に 53 個のルールを追加し全ルール数を 78 個とした場合の正判別クラスター中心型 (CC) と混合判別クラスター中心型 (MC) が，バーチャルデータ発生手法の中で最も優れた手法であるといえる．

次に，バーチャルデータの発生個数によって識別率がどのように変化するかを議論した．平均識別率が 90.00% 以上であるのは，正判別クラスター中心型 (CC) の 90.29% と混合判別クラスター中心型 (MC) の 90.18% であった．一例として，直角

表 5　Results of t-Test and Tukey's HSD Test between 25 Rules and 78 Rules

ルール形式 / 検定方法	台形型 MF・25 ルール														
	t 検定					Tukey's HSD 検定					Tukey's HSD 検定・2 fold cross validation				
バーチャルデータ発生方法	正・全 (CA)	正・ク (CC)	誤 (E)	混・全 (MA)	混・ク (MC)	正・全 (CA)	正・ク (CC)	誤 (E)	混・全 (MA)	混・ク (MC)	正・全 (CA)	正・ク (CC)	誤 (E)	混・全 (MA)	混・ク (MC)
台形型 MF 78 ルール	① ②	① ②	① ②	① ②	① ②	① ②	① ②	① ②	① ②	① ②	◯	◯	◯	◯	◯
直角台形型 MF 78 ルール	① ②	① ②	① ②	① ②	① ②	① ②	① ②	① ②	① ②	① ②	◯	◯	◯	◯	◯

台形型メンバシップ関数のルール数が 78 個における正判別クラスター中心型 (CC) の評価指標の重み (0.2, 0.5, 0.3) と混合判別クラスター中心型 (MC) の評価指標の重み (0.5, 0.25, 0.25) の 2 つの条件について，バーチャルデータの発生個数が変化したときの識別率を議論する．識別率の変化を図 7 と図 8 に示す．

図 7 は，正判別クラスター中心型 (CC) の評価指標の重み (0.2, 0.5, 0.3) の条件において，バーチャルデータの発生個数を 1 個から 10 個まで変化したときの 2-fold cross validation の 2 つのデータ集合に対するそれぞれの識別率と識別率の平均値を示している．また，図 8 は，混合判別クラスター中心型 (MC) の評価指標の重み (0.5, 0.25, 0.25) の条件において，同様に，バーチャルデータ発生個数に対する 3 種類の識別率の変化を示している．なお，バーチャルデータの発生個数を 10 個までとしたのは，正判別クラスター中心型 (CC) も混合判別クラスター中心型 (MC) も，バーチャルデータの発生個数が 1 個の場合に，識別率は 90.35% であり，バーチャルデータの発生個数が 10 個の場合には，バーチャルデータの総数が学習データの 200 個と同程度となるので，十分なバーチャルデータが発生していると考えられるからである．

図 7 では，2-fold cross validation の 2 種類のデータ集合の識別率は，バーチャルデータの発生個数の増加に対して，交差せずに一定の差を保ちながら緩やかに減少し，バーチャルデータの発生個数の変化に対する分散値は小さい．一方，識別率の平均値も，バーチャルデータが 2 個の場合の最大識別率の 90.53% をピークとして，バーチャルデータの発生個数の増加とともに緩やかに減少し，バーチャルデータの発生個数の変化に対する分散値は小さい．正判別クラスター中心型 (CC) のバーチャルデータは，誤判別データに依存せず正判別データが多いクラスター中心付近で発生するので，この特性からも，識別率は，バーチャルデータの発生個数にあまり影響されず，比較的同じような値を示すことがわかる．

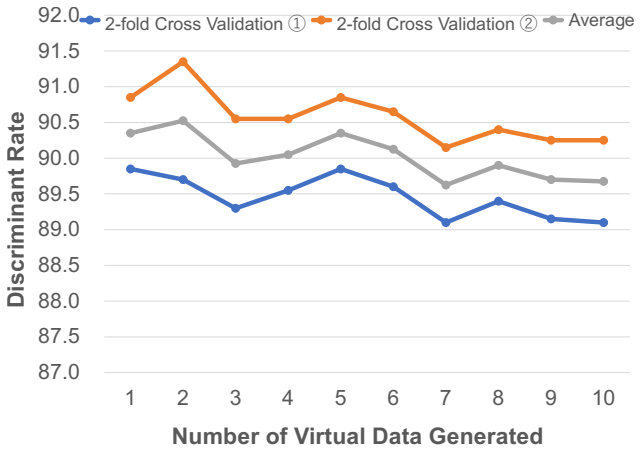


図 7　Discriminant Rates Due to Changes in Virtual Data in CC

一方，図 8 では，3 種類の識別率は，バーチャルデータの発生個数の増加とともに急激に減少している．2-fold cross validation の 2 種類のデータ集合の識別率は，バーチャルデータの発生個数の増加に対して，頻繁に交差して急激に低下し，バーチャルデータの発生個数の変化に対する分散値は大きい．一方，識別率の平均値も，バーチャルデータが 1 個の場合の最大識別率の 90.35% をピークとして，バーチャルデータの発生個数の増加とともに急激に減少し，最小識別率はバーチャルデータの発生個数が 9 個の場合の 87.90% であった．その識別率の差は 2.45% である．混合判別クラスター中心型 (MC) のバーチャルデータは，正判別データだけでなく誤判別データの付近でも発生するので，この特性からも，識別率は，バーチャルデータの発生個数に強く影響され，最適なバーチャルデータの発生個数が存在するといえる．

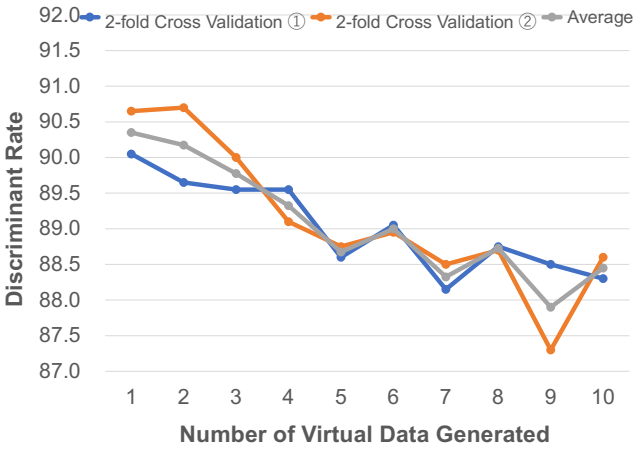


図 8　Discriminant Rates Due to Changes in Virtual Data in MC

表 6　Number of Changing Class Label per a Layer

バーチャルデータ発生方法	台形型 MF 25 ルール			台形型 MF 78 ルール			直角台形型 MF 78 ルール		
	1	5	平均	1	5	平均	1	5	平均
正・全 (CA)	0.44	1.29	0.86	0.81	3.19	2.00	0.87	2.88	1.87
正・ク (CC)	0.52	2.03	1.27	0.67	3.07	1.87	0.64	3.03	1.84
誤 (E)	3.56	19.99	11.78	2.96	18.40	10.68	2.87	17.72	10.30
混・全 (MA)	4.20	25.09	14.64	3.18	22.16	12.67	3.27	21.94	12.61
混・ク (MC)	2.06	10.92	6.49	1.69	8.66	5.17	1.59	8.27	4.93
平均	2.16	11.86	7.01	1.86	11.10	6.48	1.85	10.77	6.31

表 7　Number of Changing Class Label per a Virtual Datum

バーチャルデータ発生方法	台形型 MF 25 ルール			台形型 MF 78 ルール			直角台形型 MF 78 ルール		
	1	5	平均	1	5	平均	1	5	平均
正・全 (CA)	0.003	0.002	0.002	0.005	0.004	0.004	0.005	0.003	0.004
正・ク (CC)	0.018	0.014	0.016	0.033	0.029	0.031	0.033	0.029	0.031
誤 (E)	0.122	0.112	0.117	0.145	0.120	0.132	0.147	0.129	0.138
混・全 (MA)	0.146	0.138	0.142	0.152	0.152	0.152	0.162	0.154	0.158
混・ク (MC)	0.073	0.064	0.068	0.083	0.064	0.074	0.081	0.061	0.071
平均	0.072	0.066	0.069	0.084	0.074	0.079	0.086	0.075	0.080

最後に，クラス推定のアルゴリズムによって，バーチャルデータのクラスがどのように変化したかを議論した．1 層あたりのクラス変更回数を表 6 に示し，バーチャルデータ 1 個あた

りのクラス変更回数を表 7 に示す．ただし，変更回数は 7 種類の評価指標の重みの平均値であり，10 試行の平均値である．表 6 と表 7 から，ルール数が 25 個や 78 個でも，また，台形型や直角台形型でも，誤判別型 (E) や混合判別全領域型 (MA) のクラス変更回数は多く混合判別クラスター中心型 (MC) のクラス変更回数はそれに次いで大きい．誤判別型 (E) のバーチャルデータは判別線付近で多く発生しクラス変更の可能性が高い．同様に，混合判別全領域型 (MA) にも誤判別型が含まれ，例えば，第 l 番目の弱判別器で判別線付近の誤判別データから発生したバーチャルデータは，第 $l+1$ 番目の弱判別器に学習データとして用いられ，第 $l+1$ 番目のバーチャルデータは正判別データから発生する．このように，バーチャルデータは誤判別と正判別の両データから発生し，その発生範囲は広範囲である．その結果，クラス変更の可能性はより高くなる．混合判別クラスター中心型 (MC) でも誤判別型は含まれているが，正判別データのバーチャルデータはクラスターの中心付近で発生し発生範囲は限定される．したがって，誤判別型 (E) や混合判別全領域型 (MA) よりもクラスが変更しにくいと考えられる．一方，正判別全領域型 (CA) や正判別クラスター中心型 (CC) のクラス変更回数は少ない．これらの手法では，誤判別型が含まれていないのでバーチャルデータは判別線の付近では発生せず，クラスが変更しにくいと考えられる．また，表 7 から，バーチャルデータの発生個数の増加に対して，クラス変更回数は減少している．また，ルール数の増加に対して，クラス変更回数はわずかながら増加している．しかし，クラス変更回数はほぼ同じ規則を保っている．このことから，クラス変更回数は，バーチャルデータの発生方法に強く依存し，ルール数やバーチャルデータの発生個数にはあまり依存しないことがわかる．

これらの結果をまとめると，まず，表 1 〜表 5 と図 4〜図 6 から，識別率の高い手法は，特定領域に直角台形型メンバシップ関数を用いた 78 個ルールの正判別クラスター中心型 (CC) と混合判別クラスター中心型 (MC) であった．これらの 2 手法では，特異点データが多く存在する特定領域へのルールの追加により全体の識別率が押し上げられ，メンバシップ関数が直角台形型であるので特定領域が拡大されず，集中的にメンバシップ関数が学習され，全体の識別率に良い影響を与えたと考えられる．次に，図 7 と図 8 のバーチャルデータの発生個数と識別率の関係では，正判別クラスター中心型 (CC) は，バーチャルデータがクラスター中心付近で多い正判別データから発生するので，識別率は，バーチャルデータの発生個数に影響されず比較的一定の値を示す．混合判別クラスター中心型 (MC) は，バーチャルデータがクラスター中心付近の正判別データだけでなく全領域の誤判別データからも発生するので，識別率は，バーチャルデータの発生個数に依存する．したがって，評価指標の重みを決定した後，バーチャルデータの発生個数をパラメータとして，正判別クラスター中心型 (CC) と混合判別クラスター中心型 (MC) から最大識別率を得ることができる．ここでは，最大識別率は，バーチャルデータの発生個数が 2 個の正判別クラスター中心型 (CC) で 90.53% となった．最後に，表 6 と表 7 から，誤判別型 (E) や混合判別全領域型 (MA) は，全領域の誤判別データからバーチャルデータが発生し，クラス変更回数が多くなることから，クラス変更回数は，ルール数やバーチャルデータの発生個数に依存せず，バーチャルデータの発生方法に依存することがわかる．

5. おわりに

本論文では，pdi-Bagging のバーチャルデータの発生方法とその発生クラスを推定する手法について議論し，数値例から，バーチャルデータの発生方法とクラス変更の特性を明らかにした．今後，クラス間でデータ量に偏りが存在する場合のバーチャルデータの発生方法や Boosting のバーチャルデータの発生方法，方向性をもつデータ発生方法等を検討し，様々な質と量の計測データを用いて，実応用での有用性を検証する必要がある．

なお，本研究の一部は，JST 次世代研究者挑戦的研究プログラム (JPMJSP2150) の支援を得た．また，JSPS 科学研究費補助金基盤研究 (C) 一般「バーチャルデータ発生型ファジィバギングを用いた放送映像の卓球戦略獲得ボードの開発」(No.20K11981, 2020 年〜2025 年)，関西大学研究拠点形成支援経費「Harmonized Fitness：音楽運動のアンサンブルによる健康づくりのスマート化」(2021 年〜2022 年)，及び，2021 年度関西大学学術研究員研究費の助成を得た．

参 考 文 献

[1] 上田：アンサンブル学習，情報処理学会論文誌，Vol.46, No.SIG15(CVIM12)，pp.11-20 (2005)
[2] 村田，金森，竹ノ内：ブースティングと学習アルゴリズム：三人寄れば文殊の知恵は本当か?，電子情報通信学会誌，Vol.88, No.9，PP.724-729 (2005)
[3] R.Polikar: Ensemble Based Systems in Decision Making, *IEEE Circuits and Systems Magazine*, Vol.6, No.3, pp.21–45 (2006).
[4] L.Rokach: Taxonomy for Characterizing Ensemble Methods in Classification Tasks: A Review and Annotated ibliography, *Computational Statistics & Data Analysis*, Vol.53, No.12, pp.4046-4072, DOI:10.1016/j.csda.2009.07.017 (2009).
[5] P.Yang, Y.H.Yang, B.B.Zhou, A.Y.Zomaya: A Review of Ensemble Methods in Bioinformatics, *Current Bioinformatics*, Vol.5, No.4, pp.296-308, DOI:10.2174/157489310794072508 (2010).
[6] L.Breiman: Bagging Predictors, *Machine Learning*, Vol.24, No.2, pp.123-140 (1996).
[7] 波部：ランダムフォレスト，情報処理学会研究報告，Vol.2012-CVIM-182, No.31, pp.1-8 (2012)
[8] L.Breiman: Random Forests, *Machine Learning*, Vol.45, No.1, pp.5-32 (2001).
[9] T.G.Dietterich, G.Bakiri: Solving Multiclass Learning Problems via Error-Correcting Output Codes, *Journal of Artificial Intelligence Research*, Vol.2, pp.263-286 (1995).
[10] 仲田，鈴木：バギングによる非線形予測のリスク評価，電子情報通信学会技術研究報告. *NLP2011-60*，*CAS2011-33*，Vol.111, No.243, pp.1-6 (2011)
[11] 近藤，菊地，堀田，渋谷，前田：ランダムな特徴選択とバギングを利用した欠陥分類，画像電子学会誌，Vol.38，No.1，pp.9-15 (2009)
[12] 高橋：クラス所属確率を用いた事例ごとの分類器選択，言語処理学会第 *15* 回年次大会発表論文集，pp.709-712 (2009)
[13] A.Liaw, M.Wiener: Classification and Regression by randomForest, *The Newsletter of the R Project*, Vol.2/3, pp.18-22 (2002).
[14] Y.Freund, R.E.Schapire: A Decision-theoretic Generalization of On-line Learning and An Application to Boosting, *Journal of Computer and System Sciences*, Vol.55, No.1, pp.119-139 (1997).

[15] R.A.Jacobs, M.I.Jordan, S.J.Nowla, G.E.Hinton: Adaptive Mixtures of Local Experts, *Neural Computation*, Vol.3, pp.79–87 (1991).
[16] 村里，野口，関，上原：混合エキスパートによるブログ検索モデルの統合，情報処理学会関西支部支部大会，No.C-20 (2011)
[17] 三田：AdaBoost の基本原理と顔検出への応用，情報処理学会研究報告，Vol.42(CVIM-159), pp.265-272 (2007)
[18] 土屋，藤吉：Boosting に基づく分割統治的戦略による高精度な識別器構築手法の提案，電子情報通信学会技術研究報告 *PRMU2009-66*，Vol.109, No.182, pp.81-86 (2009)
[19] 堀内，大町，阿曽：AdaBoost アルゴリズムを用いた識別手法の統合, 電子情報通信学会論文誌, Vol.J91-D, No.4, pp.1168-1171 (2008)
[20] J.Friedman, T.Hastie, R.Tibshirani: Additive Logistic Regression: A Statistical View of Boosting, *Annals of Statistics*, Vol.28, No.2, pp.337-374 (2000).
[21] A.Torralba, K.P.Murphy, W.T.Freeman: Sharing Features: Efficient Boosting Procedures for Multiclass Object Detection, *Proceedings of IEEE Computer Society Conferenceon Computer Vision and Pattern Recognition*, pp.762-769 (2004).
[22] 石黒，澤田，坂野：多クラス早期認識ブースティング法，電子情報通信学会技術研究報告，Vol.109, No.470, pp.489-494 (2010)
[23] 池村，藤吉，森：時空間情報と距離情報を用いたマルチクラス Boosting による動作識別，電気学会論文誌 *C*，Vol.130, No.9, pp.1554-1560 (2010)
[24] 桜井，李：Joint-Boosting を用いたロバストな人物トラッキング-空間知能化のためのセンサネットワーク-，日本機械学会ロボティクス・メカトロニクス講演会講演論文集，No.2P1-N-057 (2005)
[25] 安村，上原：Bagging と Boosting を統合したアンサンブル学習方法，第 *19* 回人工知能学会全国大会予稿集，No.3F1-01 (2005)
[26] 林，鶴背：確率的データ補間を用いた BCI のための Boosting アルゴリズムの提案，信学技報，Vol.109, No.461, pp.303-308 (2010)
[27] I.Hayashi, S.Tsuruse, J.Suzuki, R.T.Kozma: A Proposal for Applying pdi-Boosting to Brain-Computer Interfaces, *Proceedings of 2012 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE2012) in 2012 IEEE World Congress on Computational Intelligence (WCCI2012)*, pp.635-640 (2012).
[28] 入江，林：正誤バーチャルデータの発生による pdi-Bagging の特性評価，第 *29* 回インテリジェント・システム・シンポジウム講演論文集，Paper ID:No.A3-3 (2019)
[29] 入江，林：pdi-Bagging による発生バーチャルデータのクラス決定法の提案，第 *34* 回人工知能学会全国大会講演論文集，Paper ID:No.103-GS-8-04 (2020)
[30] 入江，林：台形型メンバシップ関数による学習型ファジィ推論の設計評価，知能と情報，Vol.31, No.6, pp.908-917 (2019)
[31] 市橋，渡邊：簡略ファジィ推論を用いたファジィモデルによる学習型制御，日本ファジィ学会誌，Vol.2，No.3，pp.429-437 (1990)
[32] K.Kim, H.I.Choi, K.Oh: Object Detection Using Ensemble of Linear Classifiers with Fuzzy Adaptive Boosting, *EURASIP Journal on Image and Video Processing*, Vol.40, DOI;10.1186/s13640-017-0189-y (2017).
[33] J.H.Leung, Y.L.Kuo, T.W.Weng, C.L.Chin: Hybrid-Neuro-Fuzzy System and Adaboost-Classifier for Classifying Breast Calcification, *Journal of Computers*, Vol.28, No.2, pp.29-42, DOI:10.3966/199115592017042802003 (2017).
[34] A.M.Palacios, L.Sanchez, I.Couso: Using The Adaboost Algorithm For Extracting Fuzzy Rules from Low Quality Data: Some Preliminary Results, *Proceedings of 2011 IEEE International Conference on Fuzzy Systems*, pp.1263-1270 (2011).
[35] I.Lahmar, A.Zaier, M.Yahia, R.Bouallegue: A New Fuzzy Cluster Forests Method For Big Data, *Proceedings of 2019 IEEE International Conference on Internet of Things, Embedded Systems and Communications (IINTEC2019)*, pp.142-146, DOI:10.1109/IINTEC48298.2019.9112122 (2019).

［問い合わせ先］
〒569-1095 大阪府高槻市霊仙寺町 2-1-1
関西大学大学院 総合情報学研究科　林 勲
TEL: 072-690-2448
FAX: 072-690-2491
E-mail: ihaya@kansai-u.ac.jp

著 者 紹 介

入江（いりえ） 穂乃香（ほのか）［学生会員］

2019 年関西大学総合情報学部卒業，2021 年関西大学大学院総合情報学研究科博士課程前期課程修了，現在，関西大学大学院総合情報学研究科博士課程後期課程在籍，アンサンブル型クラスタリングモデルと試験管内がん細胞増殖予測に関する研究に従事．2021 年度次世代研究者挑戦的研究プログラム採択者．日本知能情報ファジィ学会の会員．

林（はやし） 勲（いさお）［正会員］

1981 年大阪府立大学工学部経営工学科卒業後，シャープ (株) 入社．1985 年大阪府立大学大学院工学研究科経営工学専攻博士前期課程修了．松下電器産業 (株)(現パナソニック (株)) 中央研究所を経て，1993 年阪南大学商学部経営情報学科講師，1997 年経営情報学部教授．2004 年より関西大学総合情報学部総合情報学科教授．1997 年南オーストラリア州立大学 KES 招聘研究員，1999 年米国ボストン大学大学院 CNS 招聘研究員，2010 年米国ボストン大学大学院 CNS 招聘教授．神経回路モデルを用いた視覚モデル，アンサンブル型クラスタリングモデル，試験管内がん細胞増殖予測，動作解析とスポーツ戦術戦略に関する研究に従事．工学博士．日本知能情報ファジィ学会第 13 期 14 期副会長，第 15 期会長．国際ファジィシステム学会 (IFSA) 理事，副会長，米国電気電子学会 (IEEE-CIS)，日本知能情報ファジィ学会，日本神経回路学会，日本視覚学会，日本基礎心理学会，システム制御情報学会等の会員．

Formulation of pdi-Bagging and Its Evaluation

by

Honoka IRIE and Isao HAYASHI

Abstract:

For pattern classification problems, there is ensemble learning method that identifies multiple weak classifiers by the learning data and combines them together to improve the discriminant rate of testing data. We have already proposed pdi-Bagging (Possibilistic Data Interpolation-Bagging) which improves the discriminant rate of testing data by adding virtually generated data to learning data. However, the accuracy of the correct virtual data type is not stable because the virtual data generate in the wide area of the data space. In addition, the discriminant accuracy is not high because the evaluation index for changing the generation class of virtual data is defined in each dimension. In this paper, we propose a new method to specify the generation area of virtual data and change the generation class of virtual data. As a result, the discriminant accuracy is improved since five new bagging methods which generate virtual data around correct discriminant data and error discriminant data are formulated, and the class of virtual data is determined with the proposed new evaluation index in multidimensional space. We formulate a new pdi-Bagging algorithm, and discuss the usefulness of the proposed method using numerical examples.

Keywords: Fuzzy Inference, Virtual Data, Ensemble Method, Bagging, Clustering

Contact Address: **Isao Hayashi, Graduate School of Informatics, Kansai University**
2-1-1, Ryozenji-cho, Takatsuki, Osaka 569-1095, Japan
TEL: +81-72-690-2448
FAX: +81-72-690-2491
E-mail: ihaya@kansai-u.ac.jp