pdi-Baggingにおける発生バーチャルデータのクラス決定法の提案

Proposal of Class Determination Method for Generated Virtual Data in pdi-Bagging

入江 穂乃香 *1　林 勲 *1
Honoka Irie　Isao Hayashi

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

We have already proposed pdi-Bagging as one of ensemble learning methods of clustering. 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 because the correct type virtual data generates near the center of class distribution with class orientation 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 this method using numerical examples.

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同定し，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学習 [1,2] に対する関心が高まっている．アンサンブル学習の一手法にバギング法 (Bagging Methods) がある．バギング法とは，データ集合からサンプリングによって複数の学習データを構成し，それらの学習データを用いて複数個の弱判別器を同定して，評価データに対する高い識別率を得る手法である．ブースティング法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，バギング法では，複数の学習データ間は独立である．我々は，弱判別器の同定の際に，仮想的に生成したデータを観測データに追加して学習データを構成し，判別線の精度を向上させる新たなバギング法やブースティング法を提案している．このバギング法を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 [3,4]．バーチャルデータの追加により，学習時にデータ量が増えるので，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する．

本論文では，バーチャルデータの発生領域を特定化し，さらに，発生クラスを変更する新たな手法を提案する．具体的には，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．正判別型では，正判別クラスのデータは判別線の近傍に位置するとは限らず，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，もしくは，特定の誤判別データから正判別データの分布性と方位性を考慮し発生領域をその分布中心付近に特定化する．また，混合型として，これらの正判別と誤判別を混合させた pdi-Bagging のアルゴリズムも定式化する．これらのバーチャルデータは観測データ集合に追加され学習データを構成する．一方，評価指標では，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義する．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 種類の評価式を加算平均して構成される．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を構成し，ファジィ推論 [5] による弱判別器で判別線を推定して，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．ここでは，これらの手法を導入した新たな 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．

連絡先: 林　勲　関西大学大学院 総合情報学研究科
〒 569-1095 大阪府高槻市霊仙寺町 2-1-1
tel. 072-690-2448
fax. 072-690-2491
e.mail　ihaya@cbii.kutc.kansai-u.ac.jp

2. pdi-Bagging

バギング法とは，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．ブースティング法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，バギング法では，複数の学習データ間は独立である．さらに，pdi-Bagging では，バーチャルデータの追加により学習データ量は増加する．

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_1 を学習し，TRD の識別率を算出する．次ステップ (層) で，メンバシップ関数を用いてバーチャルデータを特定の観測データの近傍に発生することにより TRD を増加させて，ファジィ推論の弱判別器 M_2 により TRD の識別率を算出する．TRD を増加させることによって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 $M_1, M_2, \cdots, M_i, \cdots, M_L$ に入力して，多数決により最終結果を得る．pdi-Bagging の概念図を図 1 に示す．pdi-Bagging では，バーチャルデータが学習データに追加されて弱判別器が同定されるので，従来のバギング法や AdaBoost よりも高い精度の判別線が構成される [3]．

pdi-Bagging では，弱判別器として簡易型ファジィ推論 [5]

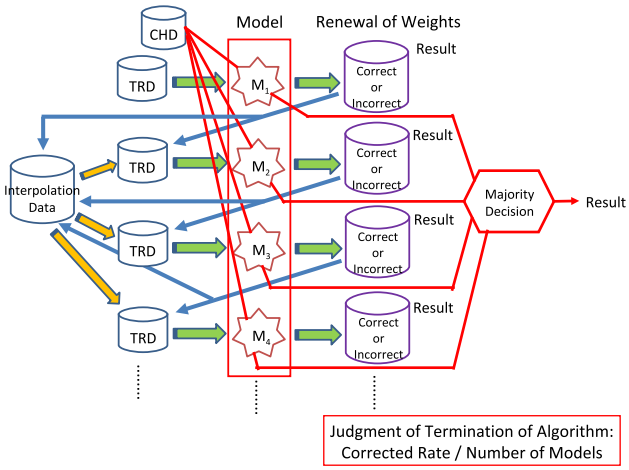

図 1: pdi-Bagging Algorithm

によるファジィクラスタリングを用いる．簡易型ファジィ推論は，if-then 型のルールをもち，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．

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

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．

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

さて，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しよう．W 個のデータからなるデータ集合 D において，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．ある特定の正判別データ $\mathbf{x}^C(d)$ や誤判別データ $\mathbf{x}^E(d)$ の周辺にバーチャルデータ $\mathbf{x}^V(d)$ を発生する．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．

正判別全領域型 ：
弱判別器により全領域での任意の $\mathbf{x}^C(d)$ のクラスが正しく判別 (正判別) された場合に，その周辺にバーチャルデータ $\mathbf{x}^V(d)$ を発生させる．

正判別クラスター中心型 ：
弱判別器により $\mathbf{x}^E(d)$ のクラスが誤判別された場合，その誤判別データに最も近い正判別データと最も遠い正判別データの中点から最近傍の正判別データ $\mathbf{x}^C(d)$ を求め，その周辺にバーチャルデータ $\mathbf{x}^V(d)$ を発生させる．

誤判別型 ：
弱判別器により $\mathbf{x}^E(d)$ のクラスが誤判別された場合に，その周辺にバーチャルデータ $\mathbf{x}^V(d)$ を発生させる．

混合型 (全領域) ：
バギングの各層ごとに，正判別全領域型と誤判別型の交互として，$\mathbf{x}^C(d)$ や $\mathbf{x}^E(d)$ の周辺にバーチャルデータ $\mathbf{x}^V(d)$ を発生させる．

混合型 (クラスター中心) ：
バギングの各層ごとに，正判別クラスター中心型と誤判別型の交互として，$\mathbf{x}^C(d)$ や $\mathbf{x}^E(d)$ の周辺にバーチャルデータ $\mathbf{x}^V(d)$ を発生させる．

なお，バーチャルデータ $\mathbf{x}^V(d)$ を構成する $x_j^V(d)$ の発生は，ある実数 h, $0 \leq h \leq 1$ 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．

$$\begin{aligned} x_j^V(d) \quad &= \quad \{x_j \mid \mu_F(x_j) = h, \ \mu_F(x_j^S(d)) = 1\} \\ h \quad &\sim \quad N(1,1), \qquad 0 \leq h \leq 1 \end{aligned}$$

ただし，$x_j^S(d)$ は $x_j^C(d)$ または $x_j^E(d)$ 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．

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

3. クラス変更化と定式化

提案した pdi-Bagging では，誤判別データや正判別データの周りに新たなバーチャルデータを発生することにより，識別率が向上する．しかし，さらに識別率を向上させるためには，バーチャルデータのクラスを正しく付け替える (変更) する必要がある．ここでは，バーチャルデータの新たなクラス決定法を提案する．

いま，バーチャルデータ $\mathbf{x}^V(d)$ が正判別データ $\mathbf{x}^C(d)$ や誤判別データ $\mathbf{x}^E(d)$ から発生したとする．新たなクラス決定法では，バーチャルデータ $\mathbf{x}^V(d)$ の新たなクラス k^* は，次の 3 つの評価基準：誤判別データの評価 (E_1)，判別クラスの評価 (E_2)，近傍クラスの評価 (E_3) から決定する．

(1) 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 $\mathbf{x}^V(d)$ とバーチャルデータを発生したクラス k をもつ正誤判別データ $\mathbf{x}^{S,k}(d)$ との距離を用いて定義する．この評価値 E_1 が小さいバーチャルデータほど，そのクラスへの依存度が高い．

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

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

(2) 判別クラスの評価 (E_2)

クラス k の中心を $\mathbf{x}_c^k$ とするとき，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 が小さいバーチャルデータほど，そのクラスへの依存度が高い．

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

(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 をもつ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 が小さいバーチャルデータほど，そのクラスへの依存度が高い．

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

これらの評価基準では，バーチャルデータが正誤判別データ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．$\mathbf{x}^V(d)$ のクラスは，評価値 E^k が最小となるクラス k^* として定義する．

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

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 から構成されるバーチャルデータを D^V で表す．

Step 2　第 i 番目の弱判別器 M_i に D^{TRD} を入力し，第 i 番目の結果 R_i の識別率 r_i^{TRD} を得る．

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．バーチャルデータ $\mathbf{x}^V(d)$ を D^V に追加する．

Step 5　結果 R_i での正判別データ数と誤判別データ数が同数になるように，乱数により D^V から $v \geq \frac{W}{2} - W^{TRD}(1 - r_t^{TRD})$ 個のバーチャルデータを取り出し D^{TRD} に加える．

Step 6　$i = i + 1$ として Step2 から 5 までを繰り返し，しきい値 θ に対して $r_i^{CHD} \geq \theta$ を満足した $K = i$ 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 $i \geq K$ を満足した時点でアルゴリズムを終了する．

Step 7　M_1, M_2, $\cdots$, M_i, $\cdots$, 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，学習データとして 100 個，評価データとして 100 個の合計 200 個である．学習データを図 2 に示す．2 入力 1 出力の 2 群判別問題として，乱数により各入力値を [0, 1] 内で発生し，2 群クラスの実数値を 2.0（赤・○印）と 3.0（青・△印）に設定した．

弱判別器には簡易型ファジィ推論を用いるので，各入力区間 [0, 1] に 5 種類のメンバシップ関数を設定した．さらに，

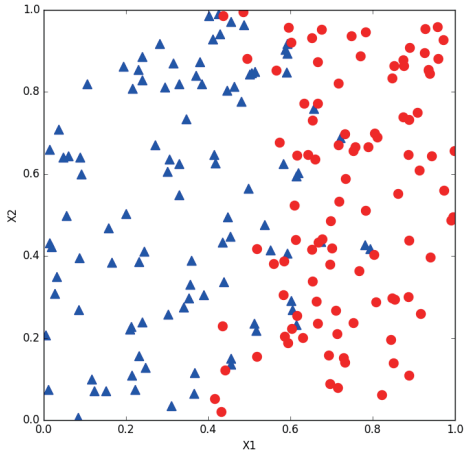

図 2: Numerical Example Data

$x_1 = 0.65$ 付近に追加のルールを設定して，ルール数は合計 196 個とした．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．各入力の 5 種類のメンバシップ関数の学習係数 K_b，K_α，K_c，K_β は同一とし 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．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 $(w_1, w_2, w_3) = \{(1/3, 1/3, 1/3), (0.2, 0.4, 0.4), (0.2, 0.3, 0.5), (0.2, 0.5, 0.3), (0.5, 0.25, 0.25)\}$ とする．バーチャルデータの発生個数は，正誤判別データから 1 個と 5 個の 2 種類とした．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．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, 10, 10), (10, 20, 20)，(10, 30, 30) の 3 種類に設定する．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，これらを 1 試行として，10 試行の正判別全領域型，正判別クラスター中心型，誤判別型，混合型 (全領域)，混合型 (クラスター中心) の識別率を比較した．ただし，混合型は，奇数層を誤判別型とし，偶数層を正判別型とする．

バーチャルデータの発生個数に対する正判別全領域型 (CA)，正判別クラスター中心型 (CC)，誤判別型 (E)，混合型 (全領域)(MA)，混合型 (クラスター中心)(MC) の学習データに対する識別率を表 1 と図 3 に示す．学習データの識別率は 10 試行の平均値である．結果から，それぞれのバーチャルデータ発生型の識別率には次の特性があることがわかる．

1) 正判別型の識別率では，クラスター中心型 (CC) が全領域型 (CA) よりも高い．全領域型 (CA) は評価指標の重みが (0.5, 0.25, 0.25) で発生個数が 1 個の場合に識別率が 90.0% と高くなったが，クラスター中心型 (CC) の識別率は全般的にそれよりも高く，最も高い識別率は 91.1% であった．

2) 誤判別型の識別率は正判別型よりも高い．評価指標の重みの変化に対して識別率の変化はなく，91.2%〜91.6% であった．

3) 混合型の識別率では，平均識別率は混合型 (全領域)(MA)

表 1: Comparison of Discrimination Rates among Five Methods

評価指標の重み	正・全 (CA)(%)			正・ク (CC)(%)			誤 (E)(%)			混・全 (MA)(%)			混・ク (MC)(%)		
	1	5	平均	1	5	平均	1	5	平均	1	5	平均	1	5	平均
(1/3, 1/3, 1/3)	89.9	89.5	89.7	90.9	90.6	90.8	91.3	91.2	91.3	91.5	91.4	91.5	91.3	92.4	91.9
(0.2, 0.4, 0.4)	89.5	89.2	89.4	90.9	90.6	90.8	91.5	91.3	91.4	91.4	91.3	91.4	91.1	92.0	91.6
(0.2, 0.3, 0.5)	89.6	89.9	89.8	90.9	90.5	90.7	91.5	91.5	91.5	91.5	90.8	91.2	91.0	91.3	91.2
(0.2, 0.5, 0.3)	89.4	89.8	89.6	90.6	90.3	90.5	91.4	91.6	91.5	91.4	91.5	91.5	91.1	91.8	91.5
(0.5, 0.25, 0.25)	90.0	89.6	89.8	91.1	90.6	90.9	91.3	91.2	91.3	91.2	91.2	91.2	91.2	91.8	91.5
平均	89.7	89.6	89.6	90.9	90.5	90.7	91.4	91.4	91.4	91.4	91.2	91.3	91.1	91.9	91.5

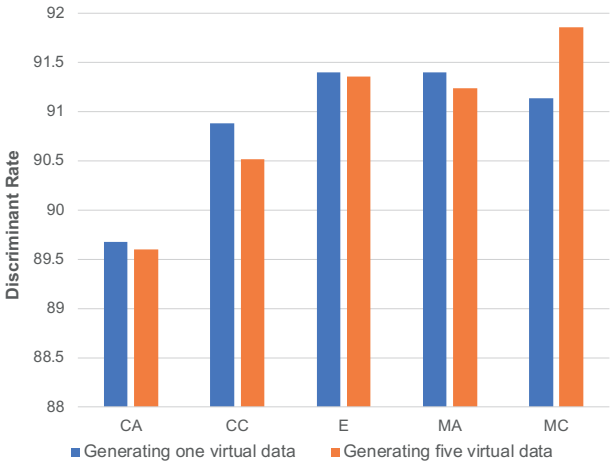

図 3: Discriminant Rate of Five Methods

も混合型 (クラスター中心)(MC) もほぼ同じであったが，混合型 (クラスター中心)(MC) で発生個数が 5 個の場合に最も高い識別率 91.9% を得た．

4) 発生個数の増減に対する識別率では，混合型 (クラスター中心)(MC) 以外は，発生個数が 1 個の方が高くなったが，混合型 (クラスター中心)(MC) では，発生個数が 5 個の場合に識別率が高く，最大識別率 92.4% を得た．

正判別型，誤判別型，混合型の比較では，正判別型は必ずしも識別率が高いとはいえないが，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) よりも高い．誤判別型の識別率は比較的良い．混合型では，混合型 (クラスター中心)(MC) の識別率は混合型 (全領域)(MA) よりも高い．バーチャルデータがクラスターの中心で発生する場合，クラスターの中心付近のファジィルールは精度良く学習され，正判別クラスター中心型 (CC) や混合型 (クラスター中心)(MC) の識別率は良くなる．一方，誤判別型はバーチャルデータが判別線付近で多く発生し，判別線付近のファジィルールが精度良く学習される．これらの結果から，正判別クラスター中心型 (CC) と誤判別型 (E) の混合型である混合型 (クラスター中心)(MC) が最も高い識別率を示したことは理解できる．

バーチャルデータの発生個数が識別率に与える影響では，誤判別型では，発生個数が 1 個でも 5 個でもほぼ同じ高い識別率を示している．判別線付近で発生するバーチャルデータは，個数が少なくても識別率に与える影響が大きいといえる．正判別型では，発生個数が識別率に与える影響はあまり大きくない．

バーチャルデータのクラス変更化では，正判別クラスター中心型 (CC) でのクラス変更回数を表 2 に示す．バーチャルデータが 1 個と 5 個の場合の変更回数を比較した．ただし，変更回数は 1 層あたりの 10 試行の平均値である．この結果から，変更回数が評価重みに依存していることがわかる．また，発生個数が 1 個の場合の平均変更回数は 0.25 回であり，5 個の場合の平均値は 1.77 回である．0.25 回の 5 倍が 1.25 回であることから，発生個数にも影響を受けていることがわかる．

表 2: Number of Changing Class Label

評価重み	1 個 (CC)	5 個 (CC)
(1/3, 1/3, 1/3)	0.04	0.55
(0.2, 0.4, 0.4)	0.32	2.12
(0.2, 0.3, 0.5)	0.12	1.92
(0.2, 0.5, 0.3)	0.64	3.42
(0.5, 0.25, 0.25)	0.12	0.82
平均	0.25	1.77

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発生クラスを変更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の特性を明らかにした．今後，様々な質と量の数値データを用いて，バーチャルデータの発生方法を検証し，本手法の実応用での有用性を検証する必要がある．

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] 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)

[4] I.Hayashi, S.Tsuruse, J.Suzuki, R.T.Kozma: A Proposal for Applying pdi-Boosting to Brain-Computer Interfaces, *Proc. of 2012 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE2012) in 2012 IEEE World Congress on Computational Intelligence (WCCI2012)*, pp.635-640 (2012).

[5] 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)