パラメータ化したFCM識別器のベンチマークテスト

市橋 秀友, 長浦一哉, 野津 亮, 本多 克宏

本研究では，FCM クラスタリング法を用いる識別器の三つの自由パラメータに加えて，さらにクラスター中心ベクトルの長さをパラメータとする場合とクラスターの混合比率をパラメータとする場合の比較を行った．提案の識別器ではマハラノビス距離による楕円状のクラスターを得るために，アルゴリズムを簡略化した繰り返し重み付最小 2 乗法に基づく更新式を用いる．ただし，自由パラメータの最適化に粒子群最適化法 (PSO) を用いるので，厳密なクラスタリングアルゴリズムの収束は必要でない．そこで，本研究では繰り返し回数を 1 回に簡略化した場合の識別性能を比較する．このことで，マハラノビス距離によるクラスタリングアルゴリズムでよく起こる発散や振動，収束までの計算時間などの問題が解消される．

クラスタリングアルゴリズムは識別器の第 1 フェーズで用いられ, FCM 識別器 (FCMC) と呼ばれる．FCM 識別器は二つのフェーズから成り, 第 1 フェーズではクラス毎にクラスタリングを行い, 第 2 フェーズでは評価用データの識別及びメンバシップ関数の自由パラメータの最適化を行う．一般に高性能識別器は, 調整できる自由パラメータを持っている．例えば, サポートベクターマシン（SVM）にはマージンやカーネルと呼ばれるパラメータがある．これらのパラメータが何らかの最適化手法で選択されることで, 識別器の汎化能力を高めている．FCM 識別器には複数の自由パラメータがあり, パラメータと誤識別率の関係は単峰形の関数ではない．そこで, パラメータ探索の簡便な手法として粒子群最適化法 (PSO) を適用する．ベンチマークデータを用いた幾通りかの分割による交差確認法（CV 法）での比較から，再代入誤識別率 (1-CV) を最小化する方法が有効であることを示す．また，自由パラメータにクラスターの混合比率，または中心ベクトルの変更割合を加えた場合の比較結果を報告する．提案 FCM 識別器は，k 最近傍法 (k-NN) よりも優れ，高性能な識別器として知られた SVM にほぼ等しい汎化性能を示した．また 10-CV 法の評価用データに対する識別精度は SVM よりも優れた結果が得られた．

キーワード：ファジィc 平均クラスタリング，識別器，粒子群最適化

1 はじめに

標準ファジィc 平均法 (FCM 法) と呼ばれるファジィクラスタリング法では，クラスター中心がデータ点に重なった場合にメンバシップ関数の分母がゼロとなり関数が不連続になることを特異であるという [1]．このことは実用上ほとんど問題にならない．しかし，標準 FCM 法のファジィ化パラメータ m を大きくしてファジィなクラスタリング結果を得ようとすると，図 1 左に示すようにメンバシップ関数がクラスター中心で尖った特異な形状になる．図の右上は 3 つのクラスターの最大メンバーシップの等値線を，右下はデータ点とクラスター中心を示している．図 2 は提案クラスタリング法（式 (2)）で $m = 1.05$，$\nu = 2$ とした場合のクラスタリング結果でガウス混合モデル [4] やエントロピー正則化 FCM 法 [2] によく似た結果となっている [2, 3]．m と ν を調節することで両者の中間的な形状にすることができる．また，ガウス混合モデルやエントロピー正則化 FCM 法では，図 3 に示すようにどちらのクラスターからも遠いデータ点 (図 3 では原点付近）のメンバシップが 1 か 0 に近いクリスプな値になってしまう性質がある (宮本 [2])．提案クラスタリング法では図 4 のようにそれらのデータ点をファジィに分類することができ，その度合いをパラメータで調節できる．これらの自由パラメータを調節して識別器の性能を高めようとするのが，提案の FCM 識別器である．

本研究では FCM 識別器の自由パラメータとして, さらにクラスター中心ベクトル v_{qj} の長さの変更割合を用いる場合とクラスターの混合比率 α_{qj} の変更割合を用いる場合の比較を行った．

提案の識別器ではマハラノビス距離による楕円状のクラスターを得るために，アルゴリズムを簡略化した繰り返し重み付最小 2 乗法 (IRLS)[5, 6] に基づく更新式を用いる [2]．ただし，クラスター中心，または混合比率の最適化に粒子群最適化法 (PSO)[7, 8, 9] を用いるので，厳密なクラスタリングアルゴリズムの収束は

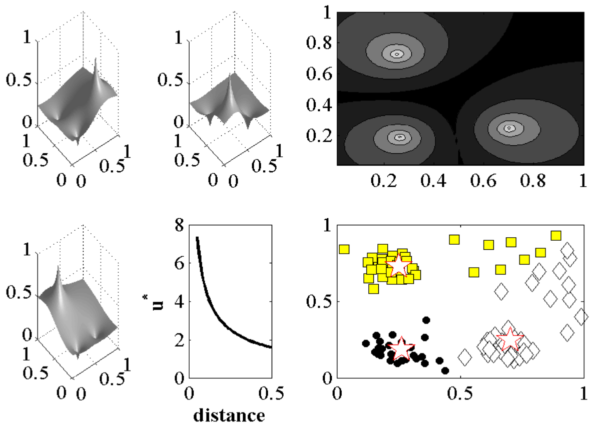

図 1: 標準 FCM 法で $m = 4$ とした場合の尖ったメンバシップ関数

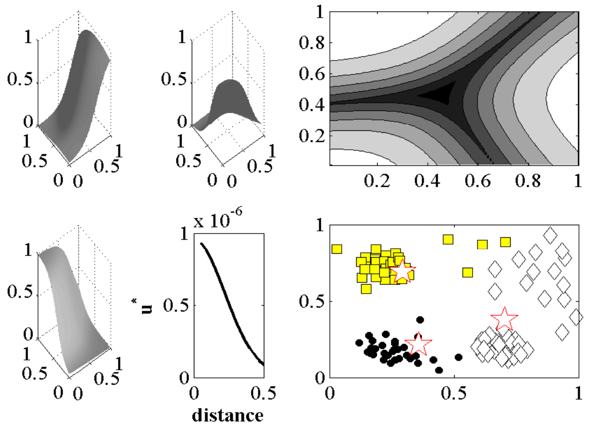

図 2: 提案クラスタリング法（式 (2)）で $m = 1.05$, $\nu = 2$ とした場合のクラスタリング結果

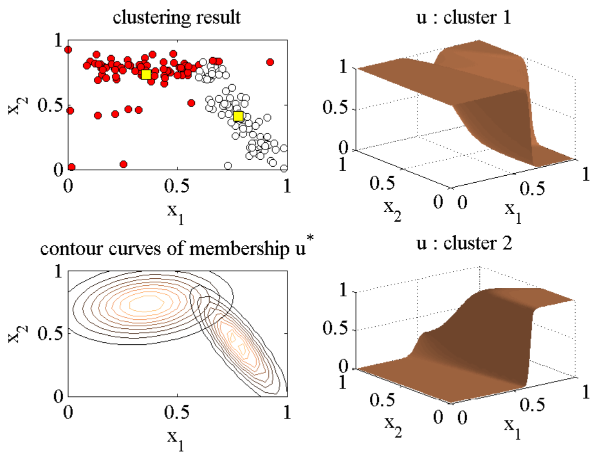

図 3: ガウス混合モデルによるクラスタリング結果

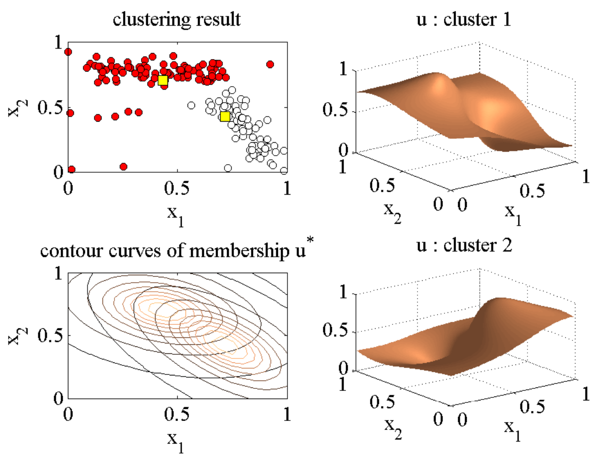

図 4: 提案クラスタリング法でパラメータの値を $m = \gamma = \nu = 1$ とした場合の結果

必要でない．そこで，本研究では繰り返し回数を 1 回に簡略化した場合の識別性能を比較する．このことで，マハラノビス距離によるクラスタリングアルゴリズムでよく起こる発散や振動，収束までの計算時間などの問題が解消される．高次元データの識別器 [10] のように厳密な収束が必要である場合のためにセミハードクラスタリング法 [11] が提案されているが本論文では用いない．

クラスタリングアルゴリズムは識別器の第 1 フェーズで用いられ, FCM 識別器 (FCMC) と呼ばれる [2, 3, 10]. FCM 識別器は 2 つのフェーズから成り, 第 1 フェーズではクラス毎にクラスタリングを行い, 第 2 フェーズでは評価用データの識別及びメンバシップ関数のパラメータ最適化を行う．一般に高性能識別器は, 調整できる自由パラメータを持っている．例えば, サポートベクターマシン (SVM)[12, 13] にはマージンやカーネルと呼ばれるパラメータがある．これらのパラメータが何らかの最適化手法で選択されることで, 識別器の汎化能力を高めている．SVM では自由パラメータが 1 個か 2 個であるのでグリッドサーチが用いられている．

FCM 識別器には複数のパラメータがあり, パラメータと誤識別率の関係は単峰形の関数ではない． そこで, パラメータ探索の簡便な手法として粒子群最適化法 (PSO) を適用する．3 章で比較するように提案識別器の PSO による最適化は単純なランダム探索に比べて大きな差は無いが，アルゴリズムの簡単さやプログラミングの容易さもランダム探索とあまり変わらない．

4 章では，ベンチマークデータを用いた幾通りかの分割による交差確認法（CV 法）での比較から，再代入誤識別率 (1-CV) を最小化する方法が有効であることを示す．また，自由パラメータにクラスターの混合比率，または中心ベクトルの変更割合を加えた場合の比較結果を報告する．提案 FCM 識別器は，k 最近傍法 (k-NN) よりも優れ，高性能な識別器として知られたサポートベクターマシン (SVM) にほぼ等しい汎化性能を示した．また 10-CV 法の評価用データに対する識別精度は SVM よりも優れた結果が得られた．

2 FCM 識別器

FCM 識別器は 2 つのフェーズに分けられる．第 1 フェーズではクラス毎にクラスタリングを行う．標準 FCM 法 [1] の目的関数は

$$J_{\mathrm{fcm}}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ki})^m D(x_k, v_i) \quad (m > 1) \quad (1)$$

と表される．ただし，$D(x_k, v_i)$ はデータベクトル $x_k \in \mathcal{R}^p$ からクラスター中心 $v_i \in \mathcal{R}^p$ までのユークリッド

距離の 2 乗で，メンバシップの m 乗 $(u_{ki})^m$ が誤差 $D(x_k, v_i)$ の重みである．これを少し一般化した FCM 法 [2, 3] の目的関数は次のように表される

$$J_{\mathrm{gfc}}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ki})^m D(x_k, v_i) + \nu\sum_{i=1}^{c}\sum_{k=1}^{N}(u_{ki})^m \tag{2}$$

最適性の必要条件（ラグランジュ関数の微分）からメンバシップの更新式は

$$u_{ki} = \left[\sum_{j=1}^{c}\left(\frac{D(x_k, v_i)+\nu}{D(x_k, v_j)+\nu}\right)^{\frac{1}{m-1}}\right]^{-1} \tag{3}$$

となる．クラスタリングアルゴリズムは v_i と u_{ki} を交互に更新し収束するまで繰り返すことになる．

式 (3) は

$$u^*_{ki} = \frac{1}{(D(x_k, v_i)+\nu)^{\frac{1}{m-1}}} \tag{4}$$

を

$$u_{ki} = \frac{u^*_{ki}}{\sum_{l=1}^{c} u^*_{kl}} \tag{5}$$

のように基準化したもので，以下の式 (11) は式 (4) から定められたものである．

楕円状のクラスターを得るための FCM 法としては Gustafson と Kessel[1, 14] による方法がよく知られている．しかし，この方法ではクラスター容量と呼ばれる S_i の行列式の値を指定する必要があるが，その値は事前には分からないという欠点がある．マハラノビス距離による楕円状のクラスターを得るために，アルゴリズムを簡略化し繰り返し重み付最小 2 乗法 (IRLS)[5, 6] に基づく更新式が用いられている [2, 3]．メンバシップ関数を重みとして最小 2 乗法 (IRLS) の目的関数は次のように定められる．

$$J_{\mathrm{ifc}}(U,V,S) = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ki}\left(D(x_k, v_i; S_i) + \log|S_i|\right) \tag{6}$$

x_k と v_i 間のマハラノビス距離を，

$$D(x_k, v_i; S_i) = (x_k - v_i)^\top S_i^{-1}(x_k - v_i) \tag{7}$$

とする．クラスター中心 v_i と第 i クラスターの分散共分散行列 $S_i \in \mathcal{R}^{p\times p}$ はラグランジュ関数の微分から次のようになる．

$$v_i = \frac{\sum_{k=1}^{N} u_{ki} x_k}{\sum_{k=1}^{N} u_{ki}} \tag{8}$$

$$S_i = \frac{\sum_{k=1}^{N} u_{ki}(x_k - v_i)(x_k - v_i)^\top}{\sum_{k=1}^{N} u_{ki}} \tag{9}$$

クラスター中心が重なり合わず競合的になるように，メンバシップ関数 u_{ki} は

$$u_{ki} = \frac{u^*_{ki}}{\sum_{l=1}^{c} u^*_{kl}} \tag{10}$$

のように基準化され, u^*_{ki} は次のように定義される．

$$u^*_{ki} = \frac{\alpha_i |S_i|^{-\frac{1}{\gamma}}}{(D(x_k, v_i; S_i)/0.1+\nu)^{\frac{1}{m}}} \tag{11}$$

u^*_{ki} は式 (4) から定めたものであるが，M 推定での Cauchy の重み関数，またはコーシー分布の確率密度関数を多次元にして，かつパラメータを増やした関数でもある．パラメータ γ は K-L 情報量正則化 FCM 法 [2] ではファジィ化パラメータとして扱われているもので，ベンチマークデータによっては有効な場合があるのでパラメータに含めた．分母の定数 0.1 はスケーリングファクターで，三つのパラメータ m，γ，ν をすべて 1 とした時に図 4 の様にファジィにクラスタリングされるように選んだ．α_i は第 i クラスターの混合比率であり，

$$\alpha_i = \frac{\sum_{k=1}^{N} u_{ki}}{\sum_{j=1}^{c}\sum_{k=1}^{N} u_{kj}} = \frac{1}{N}\sum_{k=1}^{N} u_{ki} \tag{12}$$

となる．

標準的なアルゴリズムでは $S_i, v_i, \alpha_i, u_{ki}$ を繰り返し計算するが, 本研究では繰り返し数を 1 回に設定する．これは, 第 2 フェーズでクラスターの中心や混合比率を PSO で最適化するために，クラスタリングの収束があまり意味を持たないことと, クリスプに近いクラスタリング結果を得ようとすると発散してしまう場合があるためである．クリスプに近いクラスタリング結果が誤識別率を小さくするベンチマークデータもあり，その際に特徴量（変数）の次元に対してクラスター内のデータ数が十分でないと共分散行列が特異行列となる．繰り返しアルゴリズムの確実な収束が必要な場合のためにセミハードクラスタリング [11] が提案されているが，本研究ではクラスタリングの計算時間を短くすることを優先しているので用いていない．

全てのクラス毎のクラスタリングが終わると, 第 2 フェーズとして各クラスへのメンバシップ値を求めて識別を行う．π_q をクラス q の混合比率, すなわちクラス q の事前確率とする．x_k のクラス q へのメンバシップは次のように計算される．

$$u^*_{qjk} = \frac{\alpha_{qj} |S_{qj}|^{-\frac{1}{\gamma}}}{(D(x_k, v_{qj}; S_{qj})/0.1+\nu)^{\frac{1}{m}}} \tag{13}$$

$$\tilde{u}_{qk} = \frac{\pi_q \sum_{j=1}^{c} u^*_{qjk}}{\sum_{s=1}^{Q} \pi_s \sum_{j=1}^{c} u^*_{sjk}} \tag{14}$$

c はクラス毎のクラスター数であり, Q はクラス数である．

S_i が特異行列にならないように確率的主成分分析における共分散行列の低階数近似法 [15, 16] を用いる．式 (15) の S_i' は式 (9) の S_i の近似行列である．P_i^r は $r(< p-1)$ 個の大きな固有値 $\delta_{il}, l=1,...,r$ に対応する r 個の固有ベクトルの $p \times r$ 行列を示している．p は入力データの次元数と同じである．Δ_i^r は r 個の大きな固有値 $\delta_{il}, l=1,...,r$ の $r \times r$ 対角行列である．r は全ての S_i' が特異でなく，評価用データに対する識別性能も良くなるように選択する．

S_i' の逆行列は次のようになる．

$$S_i'^{-1} = P_i^r((\Delta_i^r)^{-1} - \sigma_i^{-2} I_r) P_i^{r\top} + \sigma_i^{-2} I_p \tag{15}$$

ただし，I_r, I_p はそれぞれ r と p 次元の単位行列である．

$$\sigma_i^2 = (\mathrm{trace}(S_i) - \Sigma_{l=1}^r \delta_{il})/(p-r) \tag{16}$$

r=0 のとき，S_i' は単位行列の定数倍になる．

一般に高性能の識別器には自由パラメータが設定されていることが多い．例えば，SVM[12, 13] にはマージンパラメータやカーネルパラメータがある．交差確認法（CV 法）を用いて最も良いパラメータを決定した後に，識別器のパラメータが固定され，全てのデータで訓練される．提案の FCM 識別器では，クラス毎の全てのデータがクラスタリングされる．そして，本稼動で未知のデータに適用される．そのため識別器の性能が乱数による初期値に大きく依存するなら，最終の訓練結果は必ずしも平均的な誤識別率を保証しない．これは重要な問題である．そこで，クラスタリング結果が初期値に大きく依存せず決定論的であるように，主成分ベクトルに基づいて初期クラスター中心を決定する．4 章のベンチマークテストの結果から多くの場合にクラスター数は 1〜2 で良いと考えられる．ファジィ化パラメータを大きくした場合は事実上 1 クラスターの場合に等しい．クラスター数を多くすると $||v_{qj}||$ や α_{qj} に対応するパラメータ数が増えて最適化が難しくなる．

次式の $p_1^* \in R^p$ は一つのクラスのデータ行列 $D = (x_1, ..., x_n)^\top$ の主成分ベクトルであり，共分散行列の最大固有値，すなわち第 1 主成分（スコアー）の標準偏差 σ_1^* に対応する固有ベクトルである．本研究ではクラス毎に二つのクラスター中心とするが，その初期値は次のように与えられる．

$$v_1 = v^* + \sigma_1^* p_1^*$$

$$v_2 = v^* - \sigma_1^* p_1^* \tag{17}$$

v^* はクラスの平均ベクトルである．正規分布 $\mathcal{N}(\mu, \sigma^2)$ の場合 $\mu \pm 2\sigma$ の外側の確率は 5%，$\mu \pm 3\sigma$ の外側は 0.3% であるので，このように初期中心を決める．以下の交差確認法では，複数回（10-CV 法では 10 回）この計算が行われるので，クラス平均から第 1 クラスターの中心までのベクトルが毎回ほぼ同じ方向であるように，すなわち，一回目と二回目以降の p_1^* の内積が正になるように p_1^* のサインを決定する．

10-CV 法ではベンチマークデータを 10 個のサブセットに分ける．続いて，訓練セット（9 つのサブセット全体）から得られたクラスタリング結果を利用して残りの 1 つの評価用セットをテストする．このようにすれば全てのデータに対する誤識別件数がカウントされるので，交差確認法の誤識別率は間違って識別されたデータの全データに対する割合となる．本研究ではデータ 3 分割法（Three-way data splits, 3-WDS) で識別器を評価する．パラメータ (m，γ，ν，α_{qj}, $||v_{qj}||$) は 10-CV 法の評価用セットに対する誤識別率が最小化されるように選ばれるが，性能評価は 10-CV 法で用いられていないテストセットに対して行われる (図 5)．そして全データの 1/10 の互いに素なテストセット 10 個について繰り返されその誤識別率の平均が 3-WDS での汎化性能の評価値となる．

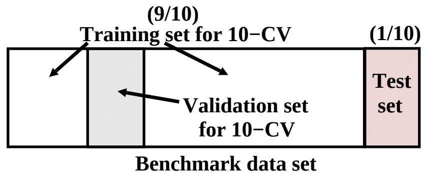

図 5: データ 3 分割法 (3-WDS)

3　PSO によるパラメータ選択

PSO とは魚や鳥の群れのように, 群を構成する個体 (粒子) 間で情報を共有しながら最良解を探す確率的な最適化手法である．粒子と呼ばれる個体は多次元空間での位置を表し, これに速度を加えることで解の探索を行う．個体の位置をベクトル $Para$ とし, 各個体の最良の位置を $pbest$, 群全体の最良の位置を $gbset$, 速度ベクトルを $Velo$ とすると, 更新則は

$$Para^{t+1} = Para^t + Velo^{t+1} \tag{18}$$

$$Velo^{t+1} = w_0 Velo^t + c_1 Rand_1(pbest - Para^t) + c_2 Rand_2(gbest - Para^t) \tag{19}$$

となる．$Para$ は最適化されるパラメータベクトルであり, m, γ, ν, α_{qj} や $||v_{qj}||$ がその要素となる．$Rand$ は区間 [0，1] の乱数の対角行列で, w_0，c_1，c_2 はスカラーの定数である．

クラスター中心を摂動させる方法の概念図を図 6 に示す．クラスターの中心ベクトル（原点は全データの重心 CG）の長さが PSO によって最適化される．CV 法においては全分割（10-CV 法の場合は 10 分割された互いに素な評価用データの集合）に対してパラメータの値は同じでなければならないため，v_{qj} をそのままパラメータ化することはできない．そこで，その長さの変更割合を自由パタメータとし，粒子の位置ベクトルの要素とする．図の黒い丸はクラスター中心であり，-0.21 はクラスター中心ベクトル（原点は全データの重心 CG）の長さを 21% 短くすることを，+0.13 は 13% 長くすることを表していて，これらの数値が PSO で最適化される．

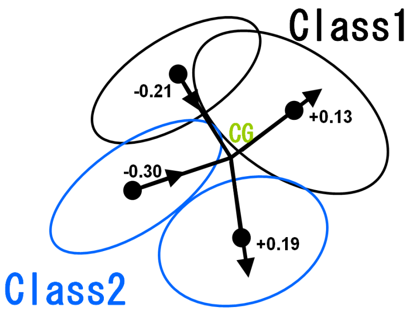

図 6: PSO によるクラスター中心の摂動

同様にして混合比率 α_{qi} の変更率をパラメータとして最適化することもできる．

4 章ではクラスター中心ベクトルの長さ $||v_{qj}||$ または α_{qj} の変更率をパラメータとして最適化する場合の比較を行う．どちらも同時に自由パラメータとすることは可能であるが，パラメータ数が増え最適化が困難になるのと，クラスの境界を摂動するのは同様であると考えられるのでどちらか一方を用いることとした．

ベンチマークデータは表 1 に示す Iris plant, Wisconsin breast cancer, Ionosphere, Glass, Liver disorder, Pima Indian diabetes, Sonar, Wine の八つのデータを用いた．これらのベンチマークデータは UCI repository[17] のものから，種々の平均をプロトタイプとする識別器の性能比較のために文献 [18] で使われたものを取り上げた．Breast cancer で欠測値のあるデータは取り除かれている．全てのカテゴリー属性は整数化し，さらに属性値は平均 0，分散 1 に基準化した．

図 7 は，Iris データでの 10-CV 法の評価用データに対する誤識別率を自由パラメータ m と γ の関数として表したもので，図 8 は Ionosphere データについてグラフにしたものである．これらのデータは 4 章で用いる UCI のベンチマークデータ（表 1）から選んだ．PSO はこれらの多峰性関数の最小値に対応するパラメータの値を探索する．Iris データでは広い範囲で平坦になっているので勾配が考慮される PSO よりもむしろランダム探索の方が有効である例を示している．Ionosphere データでは勾配の情報が有効と考えられる．Iris データのように平坦な部分の多い関数に対応するために，PSO 自体のパラメータ w_0, c_1, c_2 は表 2 のように大きな値とした．表 3 はデータ 3 分割法で 10-CV 法を 100 回実行したときの誤識別率を示している．(20×100) の場合とランダム探索である (1×2000) の場合で評価用データの誤識別 (validation error) が小さい方を太字で示している．合計の評価回数を同じにして，PSO の繰り返し回数と粒子数を変えた場合の比較である．繰り返し数 (Iteration) を 1 とした時はランダム探索である．各 10-CV 法の実行毎に PSO でパラメータ探索が行われる．

PSO の繰り返し数を 20 から 50 に増やすと，ほとん

表 1: ベンチマークデータ

	features (p)	objects (N)	classes (Q)
Iris	4	150	3
Breast	9	683	2
Iono	33	351	2
Glass	9	214	6
Liver	6	345	2
Pima	8	768	2
Sonar	60	208	2
Wine	13	178	3

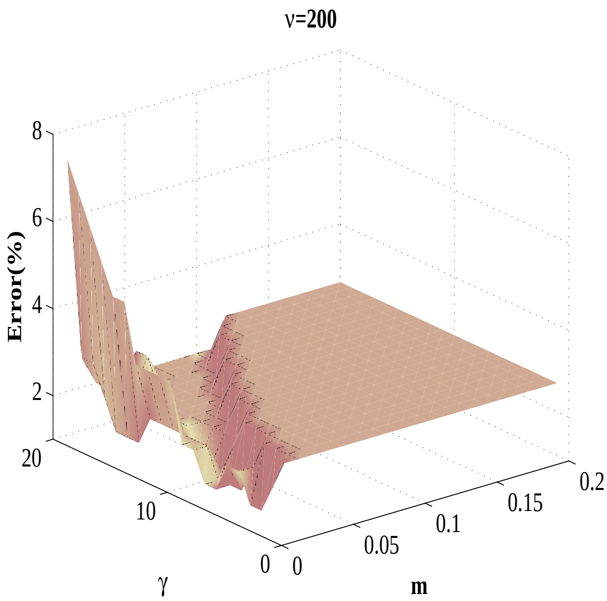

図 7: Iris データでの FCM 識別器の評価用データに対する誤識別率（m と γ の関数）

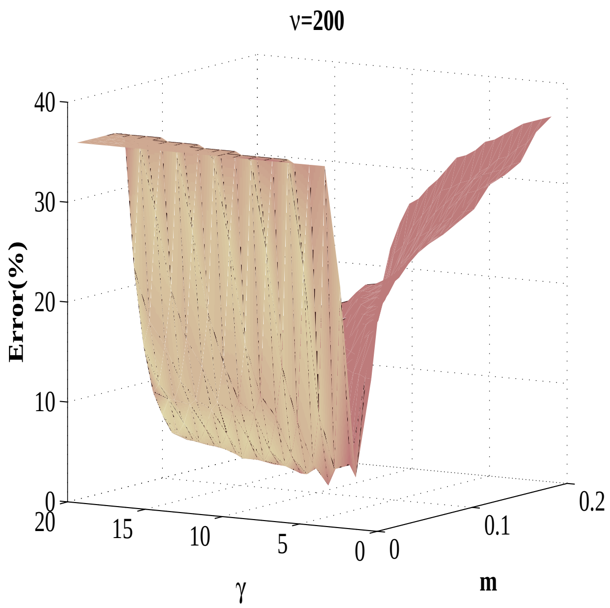

図 8: Ionosphere データでの FCM 識別器の評価用データに対する誤識別率（m と γ の関数）

表 2: PSO のパラメータ値

number of particles	100
number of iterations	20
w_0	0.8
c_1	0.4
c_2	0.4

どのデータで評価用データの誤識別率は少し良くなっている．また，20 の場合は五つのデータ（Iono，Glass，Liver，Pima，Sonar）でランダム探索よりも良くなっている．改善される度合いは僅かであるがテスト用データについても五つのデータで同じ傾向が見られる．Breast Cancer データではほとんど同じで，Wine データではほとんどゼロになっている．ランダム探索による Iris データの結果は繰り返し数が 20 の場合より良くなっている．これらの八つのデータ中で Iris だけが例外的であるが，図 7 の平坦な領域の広いグラフから分かるように，このような場合にはランダムネスを大きくする必要がある．また，繰り返し数を増やした場合にテスト用データに対する誤識別は増えていない．これらの値は計算時間と誤識別率の改善度合いで決定すべきである．表 3 の結果から次章の性能比較では，PSO における繰り返し数は 20，粒子数は 100 とした．

4　ベンチマークデータの識別結果

PSO によるパラメーター探索は，CV 法での評価用データの誤識別率を最小化するように行われる．その

表 3: データ 3 分割法（100 回）における PSO の繰り返し数と粒子数の誤識別率 (%) への影響の比較

	iteration×particls	validation error	test error
Iris	20×100	2.00±0.10	3.27±0.66
	1×2000	**1.84**±0.07	2.40±0.56
	50×100	1.96±0.07	3.13±0.63
Breast	20×100	2.51±0.01	3.15±0.21
	1×2000	2.51±0.01	2.97±0.12
	50×100	2.51±0.01	2.94±0.23
Iono	**20×100**	**3.61**±0.15	4.17±0.39
	1×2000	4.06±0.09	4.69±0.39
	50×100	3.37±0.13	4.00±0.52
Glass	**20×100**	**26.98**±0.23	29.81±1.06
	1×2000	27.87±0.19	30.19±0.85
	50×100	26.31±0.25	29.81±1.53
Liver	**20×100**	**25.50**±0.16	28.71±1.14
	1×2000	26.40±0.18	29.06±0.84
	50×100	25.22±0.13	28.21±0.79
Pima	**20×100**	**22.98**±0.05	24.64±0.52
	1×2000	23.34±0.07	25.01±0.45
	50×100	22.76±0.09	24.89±0.51
Sonar	**20×100**	**11.07**±0.13	14.40±1.39
	1×2000	12.06±0.13	15.65±1.86
	50×100	10.78±0.20	14.90±1.17
Wine	20×100	0.04±0.04	0.06±0.19
	1×2000	0.03±0.03	0±0
	50×100	0.04±0.05	0.12±0.25

ため，CV 法の評価用データの誤識別率は未知のデータに対する誤識別率より小さくなると考えられる．そこで，FCM 識別器の汎化性能を正確に測定するためにデータ三分割法 (3-WDS) を適用する．3-WDS はデータを訓練データ，評価用データ，テスト用データに分割する手法である．訓練データはクラスタリングにより，クラスターの中心と分散共分散行列を求めるために使用される．評価用データは CV 法によって識別器のパラメーターを調節するために使用される．テスト用データは未知のデータとして扱われ，パラメータの調整が終了した後で識別器の性能評価に用いられる．

本章のテストでは，第一フェーズのクラスタリングは $m = 0.1, \gamma = 1, \nu = 5$ として，繰り返し数を 1 とした．表 5〜表 13 は，データセットの 1/10 をテスト用データとして性能評価のみに用いた結果である．残りの 9/10 には CV 法を適用して訓練データと評価用データに分割し，PSO で評価用データの誤識別率が最小となるようにパラメータ値を選択した．そして，9/10 のデータに対するクラスタリング結果と選択されたパラメータの値を用いて，テスト用データの誤識別率を評価した．この際に 1/10 のデータの選び方で結果が変動するので，1/10 のテスト用データ 10 個が互いに素であるように分割して，以上の手順を 10 回繰り返し，全

てのデータに対するテスト用データでの誤識別率を求めた．3-WDS は 10 回で 1 つのセッションを構成しており, 1 セッションで全てのデータが 1 度ずつテストされる．“resubstitution”(1-CV, 再代入法) での訓練セットはデータの 9/10 から成り， テストセットは残りの 1/10 で 10 回で 1 セッションとする．

表 4 は CV 法の種々の分割法での m, γ, ν と $||v_{qj}||$ の変更率をパラメータとする FCMC($||v_{qj}||$) の評価用データに対する誤識別率を示している．10-CV での誤識別率は文献 [3] と [19] に掲載されている SVM での結果に比べてすべてのデータで小さくなっている．表 5 と表 6 は評価用データの分割方法を変えてテスト用データの誤識別率を測定した場合の結果である．3-WDS は 1 セッション行った．表 5 は FCMC($||v_{qj}||$) の誤識別率を示す．♠ は表 6 の k-NN の “10-CV” 欄と比べて誤識別率が大きい又は等しい場合を示している．下線付きの太字は各データセットにおける最も小さい誤識別率を示す．表 5 でテスト用データに対して最も識別性能が良かったのは，誤識別率が最小となるデータが四つで, ♠ がない “resubstitution”(1-CV, 再代入法) である．

表 4: CV 法の種々の分割法での FCMC($||v_{qj}||$) の評価用データに対する誤識別率 (%)

	resubstitution	20-CV	10-CV	5-CV
Iris	0.67	0.75	1.77	1.41
Breast	2.34	2.50	2.38	2.34
Iono	2.97	3.33	3.45	3.52
Glass	7.98	27.44	25.79	29.32
Liver	21.06	23.87	22.61	22.77
Pima	20.33	22.16	22.26	22.26
Sonar	1.86	12.00	12.00	10.86
Wine	0.00	0.06	0.06	0.44

表 6 は k-NN の誤識別率を示す．パラメータ k は 1〜50 の全ての整数値をテストし最良値を選んだ．♠ は表 5 の “resubstitution” より誤識別率が大きい又は等しいことを示している．FCMC($||v_{qj}||$) は多くのベンチマークセットに対して k-NN より優れた識別性能を示した．

表 5 と表 6 の結果より, 以下では, PSO の最適化を “resubstitution” と “10-CV” に限定して比較を行う．3-WDS のセッションを 10 回実行するので, 再代入法や 10-CV 法は 100 回実行することになる．したがって, ベンチマークデータの全てのサンプルは 10 回テストされる．

表 5: CV 法の種々の分割法での FCMC($||v_{qj}||$) のテスト用データに対する誤識別率 (%)

	resubstitution	20-CV	10-CV	5-CV
Iris	<u>**2.67**</u>	4.00	3.33	5.33
Breast	2.94	♠ 3.38	2.50	<u>**2.35**</u>
Iono	4.29	4.29	<u>**4.00**</u>	5.14
Glass	<u>**25.71**</u>	♠ 30.00	♠ 30.95	♠ 29.05
Liver	28.53	26.76	<u>**25.59**</u>	26.18
Pima	24.21	<u>**23.42**</u>	23.95	23.82
Sonar	<u>**15.00**</u>	♠ 18.00	♠ 16.50	♠ 15.50
Wine	<u>**0.59**</u>	1.18	<u>**0.59**</u>	<u>**0.59**</u>

表 6: CV 法の種々の分割法での k-NN のテスト用データに対する誤識別率 (%)

	k-NN classifier			
	resubstitution	20-CV	10-CV	5-CV
Iris	♠ 5.33	♠ **4.67**	♠ 6.00	♠ **4.67**
Breast	♠ 4.71	♠ 3.53	♠ 3.24	♠ **3.09**
Iono	♠**13.14**	♠**13.14**	♠**13.14**	♠**13.14**
Glass	♠30.48	♠28.57	♠**28.10**	♠30.48
Liver	♠**34.71**	♠35.59	♠35.00	♠36.18
Pima	♠29.74	♠**24.47**	♠**24.47**	♠25.00
Sonar	**14.00**	♠17.00	♠15.50	15.00
Wine	♠ 4.12	♠ **2.94**	♠ 4.12	♠ **2.94**

表 7 に k-NN による “resubstitution” と “10-CV” でのテスト用データに対する誤識別率を再掲し評価用データと訓練用データに対する誤識別率を追加して示す．以下ではこの表を元に FCM 識別器の性能を比較する．“Training error” は訓練データに対する誤識別率 (再代入誤識別率), “Validation error” は評価用データに対する誤識別率, “Test error” はテスト用データに対する誤識別率を示す．表 7 より, Liver の “Test error” はほぼ同等であり, “resubstitution” は二つのデータ（Iris, Sonr）で “10-CV” より誤識別率が小さく, “10-CV” は三つのデータ（Breast, Glass, Pima）で “resubstitution” より小さい結果となった．“resubstitution” のテスト用データに対する誤識別率は, “10-CV” より小さくなる傾向があるが大きな差は無い．k-NN では訓練データが全て記憶されるので, 1-NN 識別器では訓練データに対する誤識別が 0 になる．

表 8〜表 10 はパラメータ選択に 10-CV を用いた FCM 識別器の誤識別率 (%) ± 標準偏差を示す．表 11〜表 13 は再代入誤識別率を最小化するようにパラメータ選択した場合 (Trained by resubstitution, 1-CV) である．

表 8 と表 11 は m, γ, ν のみをパラメータとする FCM

識別器 (FCMC) の誤識別率を示す． 表 9, 表 12 は m, γ, ν と α_{qj} の変更率をパラメータとした FCM 識別器 (FCMC(α_{qj})) の誤識別率を示す．表 10 と表 13 は m, γ, ν と $||v_{qj}||$ の変更率をパラメータとした FCM 識別器 (FCMC($||v_{qj}||$)) の誤識別率を示す．“k” の欄は FCM 識別器と k-NN との平均値の差の t 検定の結果を示す．◯ は FCM 識別器の誤識別率が k-NN の誤識別率より両側検定 ($p < 0.05$) で有意に小さいことを表す．♠ は FCM 識別器の誤識別率が k-NN の誤識別率より優位に大きいことを表す．

表 8 では, Sonor 以外の全てのデータで FCMC の方が k-NN より評価用データの誤識別率が小さい．テスト用データの誤識別率は, Glass と Pima の場合だけ k-NN の方が小さい．

表 9 の “k” 欄と “F” 欄は, FCMC(α_{qj}) が k-NN や三つのパラメータのみの FCMC に比べて, 明らかに評価用データの識別性能が良いことを示している．テスト用データの誤識別率では，k-NN が FCMC(α_{qj}) より有意に小さいのは Glass のみである．FCMC(α_{qj}) と三つのパラメータのみの FCMC との比較では, Iris と Breast は FCMC の方が小さく, Iono, Pima, Sonar は FCMC(α_{qj}) の方が小さいので, 明白な違いは認められなかった．

表 10 では, FCMC のテスト用データの誤識別率が FCMC($||v_{qj}||$) より有意に小さいのは Iris のみである．k-NN は Glass と Sonar で FCMC($||v_{qj}||$) より小さくなっている．

表 11-13 は, テスト用データを除く全てのデータを用いて訓練と PSO によるパラメータ最適化を行った場合，すなわち再代入誤識別率を最小化するようにした場合の，それぞれ FCMC, FCMC(α_{qj}), FCMC($||v_{qj}||$) の結果を示す．表 11 より, テスト用データの識別では Pima と Sonar 以外はパラメータが三つだけの FCMC の方が 10-CV の k-NN(誤識別率は表 7, “Trained by 10-CV”-“Test error” に示したもの) より誤識別率が小さいか等しい．表 12 は FCMC(α_{qj}) の場合で，テスト用データの識別では, 全てのベンチマークデータにおいて FCMC(α_{qj}) の方が k-NN(resubstitution) より有意に誤識別率が小さいか等しくなった．k-NN が FCMC(α_{qj}) より誤識別率が小さいのは “10-CV”-“k” 欄の Glass のみである．表 7 で, 10-CV を用いた場合の k-NN の Glass は, テスト用データの誤識別率が評価用データの誤識別率より小さい. このことから, k-NN での Glass の結果は少し特殊な場合であるといえる．

表 13 は再代入誤識別率を最小化するようにした場合の FCMC($||v_{qj}||$) の誤識別率と性能比較を示す．テスト用データの識別結果は, 全てのデータセットにおいて, 10-CV を用いた k-NN より誤識別率が有意に小さいか等しくなった．表 12，表 13 の結果から，再代入法で（訓練データに対する誤識別率を最小化するように）パラメータ最適化を行った FCM 識別器は高い汎化能力を示すと言える．このことは，FCM 識別器を用いる際には全てのデータに対して訓練を行うだけでよいことを意味している．訓練データが大量にある場合は，訓練データの誤識別率が重要となる．

表 8 から表 13 の太字で示す誤識別率は，3 通りの自由パラメータの設定法（m, γ, ν のみの場合，α_{qj} の変更率を含む場合，$||v_{qj}||$ の変更率を含む場合）と 2 通りの CV 法（10-CV, 1-CV）の組み合わせの中でデータ毎に最も誤識別率が小さいものを示している．全てのデータが 10 回ずつテスト用データとして選ばれた結果であるので，もしベンチマークデータ毎にこれらの方策を選択するなら 3-WDS での性能は大きく改善される．また自由パラメータが m, γ, ν のみの FCM 識別器の場合は太字の結果が無いことから，自由パラメータを増やすことの効果は大きいと言える．

表 14, 表 15 は，それぞれ FCMC(α_{qj}), FCMC($||v_{qj}||$) の k-NN 及びサポートベクターマシン (SVM) との性能比較を示す．SVM は優れた識別手法として知られている．性能比較はランダムに選ばれた 1/3 のテスト用データを用いて行い, 識別器の訓練やパラメータ最適化は残りの 2/3 を用いて行った．訓練／評価用データはランダムに選択され, 選択されなかった 1/3 のデータをテスト用データとした．比較する三つの識別器のテスト用データに対する性能は，全て以下の手順で評価されている．

1. ランダムに選択した訓練／評価用データに 10-CV を適用して，評価用データの誤識別率が最も小さくなるパラメータ値を選択する．
2. 求めたパラメータを用いて識別器を構築する．
3. 残りの 1/3 のテスト用データに対する誤識別率を求め, 識別器の精度を評価する．

表 14, 表 15 の SVM の誤識別率と標準偏差 (“SVM-RBF 10 runs” 欄) は, ステップ 1 から 3 を 10 回行ってその平均を求めたものであり, 文献 [19] から引用した．SVM-RBF のパラメータ（kernel, regularization）は 10-CV での評価用データの誤識別を最少化するようにグリッドサーチにより求められている．最適なパラメータは訓練／評価用データの選び方で大きく変化するので, 識別性能を厳密に測定するために, FCM 識別器と k-NN についてはステップ 1 から 3 を 100 回行って誤識別率の平均を求めた．Glass は文献 [19] に

含まれていないので, Statlog Australian credit (Australia, N=690, p=14, Q=2), Statlog German credit (German, N=1000, p=20, Q=2), Statlog heart disease (Heart, N=270, p=13, Q=2) の三つのベンチマークデータを加えた．

FCM 識別器, k-NN, SVM の性能比較は平均値の差の両側 t 検定 ($p \leq 0.05$) を用いた．t 検定の結果は表の "k" 欄, "S" 欄に示す．"S" は SVM との比較結果を表す．表 14, 表 15 で ⊙ は誤識別率が 5% 以上小さいか, 半分以下であることを示している．表 14 より, FCMC(α_{qj}) は多くのデータで k-NN より誤識別率が小さい．表 15 は FCMC($||v_{qj}||$) の識別性能を示す．Pima は "SVM" と同等となっている．計算時間もここで取り上げたベンチマークデータでは大きな差は無いが, SVM ではカーネル行列のサイズが $N \times N$ で, 訓練データ数が大量であればワーキングセットに分解して計算を繰り返すこと（decomposition method）が必要で, 誤識別が多くサポートベクターが大量になる場合も計算時間が増大する．提案法ではほぼ訓練データ数に比例して計算時間が増えるので, 訓練データ数が多い場合に有効である．

以上より, FCM 識別器はテスト用データと評価用データの両者に高精度な識別器であると言える．

表 7: k-NN の誤識別率 (%)

	k-NN classifier			
	Trained by 10-CV		Trained by resubstitution	
	Validation error	Test error	Training error	Test error
Iris	3.85	6.00	0.0	5.33
Breast	3.02	3.24	0.0	4.71
Iono	13.10	13.14	0.0	13.14
Glass	29.42	28.10	0.0	30.48
Liver	32.55	35.00	0.0	34.71
Pima	23.75	24.47	0.0	29.74
Sonar	13.28	15.50	0.0	14.00
Wine	2.25	4.12	0.0	4.12

表 8: m, γ, ν のみをパラメータとする FCMC の 10-CV での誤識別率 (%)

	Trained by 10-CV			
	Validation error		Test error	
	FCMC	k	FCMC	k
Iris	2.26 ± 0.09	○	2.67 ± 0.54	○
Breast	2.90 ± 0.00	○	2.79 ± 0.00	○
Iono	4.04 ± 0.08	○	4.86 ± 0.52	○
Glass	27.36 ± 0.23	○	30.29 ± 1.29	♠
Liver	26.33 ± 0.19	○	28.38 ± 0.66	○
Pima	23.46 ± 0.04	○	25.24 ± 0.39	♠
Sonar	13.64 ± 0.13	♠	15.75 ± 0.72	-
Wine	0.31 ± 0.05	○	0.18 ± 0.28	○

表 9: 10-CV での FCMC(α_{qj}) の誤識別率 (%)

	Trained by 10-CV					
	Validation error			Test error		
	FCMC(α_{qj})	k	F	FCMC(α_{qj})	k	F
Iris	2.00 ± 0.10	○	○	3.27 ± 0.66	○	♠
Breast	2.50 ± 0.01	○	○	3.15 ± 0.21	-	♠
Iono	3.61 ± 0.14	○	○	**4.17** ± 0.39	○	○
Glass	26.98 ± 0.23	○	○	29.81 ± 1.06	♠	-
Liver	25.50 ± 0.16	○	○	28.71 ± 1.14	○	-
Pima	22.98 ± 0.05	○	○	24.64 ± 0.52	-	○
Sonar	11.07 ± 0.13	○	○	**14.40** ± 1.39	○	○
Wine	0.04 ± 0.04	○	○	**0.06** ± 0.19	○	-

表 10: 10-CV での FCMC($||v_{qj}||$) の誤識別率 (%)

	Trained by 10-CV					
	Validation error			Test error		
	FCMC($||v_{qj}||$)	k	F	FCMC($||v_{qj}||$)	k	F
Iris	1.65 ± 0.10	○	○	4.07 ± 0.86	○	♠
Breast	2.37 ± 0.02	○	○	**2.56** ± 0.10	○	○
Iono	3.64 ± 0.09	○	○	4.57 ± 0.43	○	-
Glass	26.36 ± 0.31	○	○	30.71 ± 1.25	♠	-
Liver	23.21 ± 0.23	○	○	**26.03** ± 1.14	○	○
Pima	22.36 ± 0.05	○	○	23.59 ± 0.37	○	○
Sonar	12.26 ± 0.17	○	○	16.80 ± 1.48	♠	-
Wine	0.06 ± 0.03	○	○	0.53 ± 0.59	○	-

5　おわりに

本論文では FCM 識別器の性能評価を行った．提案の FCM 識別器のテスト用データに対する汎化性能は, 高精度識別器として知られる SVM にほぼ等しく，評価用データに対してはより優れた結果が得られた．また, 訓練データに対する誤識別率，すなわち再代入誤識別率を最小化する場合も高い汎化性能が得られた．パラメータ数を増やすと最適化は煩雑になるが，PSO はランダム探索に似て簡便であり，識別器をパラメータ化して性能を向上させる際に実装が容易である．

参考文献

[1] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[2] S. Miyamoto, H. Ichihashi, and K. Honda, *Algorithms for Fuzzy Clustering, Methods in c-Means Clustering with Applications*, Springer-Verlag, Berlin, 2008.

表 11: m, γ, ν のみをパラメータとする FCMC の再代入法での誤識別率 (%)

	Trained by resubstitution (1-CV)		
	Training error	Test error	
			10-CV
	FCMC	FCMC	k
Iris	0.71 ± 0.04	3.27 ± 0.50	○
Breast	2.78 ± 0.00	2.96 ± 0.15	○
Iono	3.31 ± 0.10	4.71 ± 0.41	○
Glass	8.58 ± 0.15	27.71 ± 0.63	-
Liver	22.40 ± 0.10	28.50 ± 0.59	○
Pima	21.28 ± 0.06	25.13 ± 0.59	♠
Sonar	2.60 ± 0.05	16.35 ± 0.47	♠
Wine	0.00 ± 0.0	0.41 ± 0.40	○

表 12: 再代入法での FCMC(α_{qj}) の誤識別率 (%)

	Trained by resubstitution (1-CV)						
	Training error		Test error				
						10-CV	
	FCMC(α_{qj})	F	FCMC(α_{qj})	k	F	k	F
Iris	0.67 ± 0.0	-	3.07 ± 0.56	○	-	○	-
Breast	2.49 ± 0.01	○	3.09 ± 0.17	○	-	○	♠
Iono	2.94 ± 0.16	○	4.34 ± 0.44	○	-	○	○
Glass	8.18 ± 0.18	○	**28.90** ± 0.92	○	♠	♠	○
Liver	21.83 ± 0.09	○	28.11 ± 1.01	○	-	○	-
Pima	20.81 ± 0.05	○	24.79 ± 0.55	○	-	-	○
Sonar	2.11 ± 0.13	○	**14.40** ± 0.88	-	○	○	○
Wine	0.0 ± 0.0	-	0.89 ± 0.42	○	♠	○	♠

表 13: 再代入法での FCMC($\|v_{qj}\|$) の誤識別率 (%)

	Trained by resubstitution (1-CV)						
	Training error		Test error				
						10-CV	
	FCMC($\|v_{qj}\|$)	F	FCMC($\|v_{qj}\|$)	k	F	k	F
Iris	0.67 ± 0.00	○	**2.80** ± 0.82	○	-	○	-
Breast	2.34 ± 0.01	○	2.71 ± 0.17	○	○	○	-
Iono	3.00 ± 0.07	○	4.40 ± 0.31	○	-	○	○
Glass	8.71 ± 0.24	-	29.24 ± 1.91	-	♠	-	-
Liver	21.14 ± 0.12	○	26.82 ± 1.10	○	○	○	○
Pima	20.47 ± 0.07	○	**23.54** ± 0.69	○	○	○	○
Sonar	1.95 ± 0.06	○	15.50 ± 0.97	♠	○	-	-
Wine	0.00 ± 0.0	-	1.18 ± 0.62	○	♠	○	♠

[3] H. Ichihashi, K. Honda, A. Notsu and K. Ohta, "Fuzzy c-means classifier with particle swarm optimization," *Proc. of the IEEE International Conference on Fuzzy System, World Congress on Computational Intelligence*, Hong Kong, China, pp. 207-215, 2008.

[4] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[5] P. W. Holland and R. E. Welsch, "Robust regres-

表 14: FCMC(α_{qj}), k-NN, SVM の誤識別率 (%)

	Test set error rate trained by 10-CV				
	FCMC(α_{qj})			k-NN	SVM-RBF
	100 runs	k	S	100 runs	10 runs
Iris	3.98 ± 2.52	○	-	5.68 ± 2.91	3.4 ± 3.4
Breast	3.06 ± 0.86	○	-	3.37 ± 0.10	3.6 ± 1.0
Iono	4.64 ± 1.79	⊙	-	14.86 ± 3.59	4.6 ± 1.7
Liver	30.47 ± 3.68	⊙	-	37.94 ± 4.37	29.6 ± 3.2
Pima	24.36 ± 2.47	○	♠	25.77 ± 2.60	22.7 ± 2.2
Sonar	17.46 ± 4.30	♠	⊙	15.80 ± 4.17	25.0 ± 6.6
Wine	1.97 ± 2.05	⊙	-	4.36 ± 2.89	2.2 ± 2.1
Australia	14.34 ± 2.03	-	-	14.39 ± 2.00	13.7 ± 1.8
German	24.17 ± 2.02	○	-	27.83 ± 2.12	24.1 ± 1.4
Heart	16.39 ± 3.66	-	-	17.31 ± 3.57	15.3 ± 4.8

表 15: FCMC($\|v_{qj}\|$), k-NN, SVM の誤識別率 (%)

	Test set error rate trained by 10-CV				
	FCMC($\|v_i\|$)			k-NN	SVM-RBF
	100 runs	k	S	100 runs	10 runs
Iris	3.84 ± 2.44	○	-	5.68 ± 2.91	3.4 ± 3.4
Breast	3.03 ± 1.12	○	-	3.37 ± 0.10	3.6 ± 1.0
Iono	5.06 ± 2.11	⊙	-	14.86 ± 3.59	4.6 ± 1.7
Liver	29.11 ± 3.85	⊙	-	37.94 ± 4.37	29.6 ± 3.2
Pima	23.96 ± 2.55	○	-	25.77 ± 2.60	22.7 ± 2.2
Sonar	18.16 ± 4.84	♠	⊙	15.80 ± 4.17	25.0 ± 6.6
Wine	2.31 ± 2.49	○	-	4.36 ± 2.89	2.2 ± 2.1
Australia	13.78 ± 1.96	○	-	14.39 ± 2.00	13.7 ± 1.8
German	24.12 ± 2.37	○	-	27.83 ± 2.12	24.1 ± 1.4
Heart	16.78 ± 3.26	-	-	17.31 ± 3.57	15.3 ± 4.8

sion using iteratively reweighted least-squares," *Communications in Statistics*, vol. A6, no. 9, pp. 813-827, 1977.

[6] P. J. Huber. *Robust Statistics*. New York:Wiley, first edition, 1981.

[7] R. C. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," *Proc. of the 6th International Symposium on Micro Machine and Human Science*, Nagoya, Japan, pp. 39-43, 1995.

[8] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," *Proc of the IEEE International Conference on Neural Networks*, Piscataway, NJ, vol. 4, pp. 1942-1948, 1995.

[9] M. Clerc, *Particle Swarm Optimization*, Wiley, 2006.

[10] H. Ichihashi, K. Honda, A. Notsu and E. Miyamoto, "FCM classifier for high-dimensional data," *Proc. of 2008 IEEE International Conference on Fuzzy System, World Congress on Computational Intelligence*, Hong Kong, China, pp. 200-206, 2008.

[11] H. Ichihashi, A. Notsu and K. Honda, "Triplet of FCM classifiers," *Proc. of 2009 IEEE International Conference on Fuzzy System*, Jeju, Korea, Aug. 20-24, 2009.

[12] V.N. Vapnik, *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, Berlin, 1982.

[13] C. Cortes and V. Vapnik, "Support-vector network." *Machine Learning*, vol.20, pp. 273-297, 1995.

[14] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," *Proc. IEEE CDC*, vol.2, pp. 761-766, 1979.

[15] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol.11, pp. 443-482, 1999.

[16] F. Sun, S. Omachi, and H. Aso, "Precise selection of candidates for hand written character recognition," *IEICE Trans. Information and Systems*, vol.E79-D, no.3, pp. 510-515, 1996.

[17] A. Asuncion and D.J. Newman, UCI Repository of machine learning databases [http://www.ics.uci.edu/~ mlearn/MLRepository.html]. Irvine, CA: University of California, Dept. of Information and Computer Science, 2007.

[18] C. J. Veenman and M.J.T. Reinders, "The nearest sub-class classifier: a compromise between the nearest mean and nearest neighbor classifier," *IEEE Transactions on PAMI*, vol.27, no.9, pp. 1417-1429, 2005.

[19] T. V. Gestel et al., "Benchmarking least squares support vector machine classifiers," *Machine Learning*, vol. 54, pp. 5-32, 2004.

[問い合わせ先]
〒 599-8531 堺市中区学園町 1 - 1
大阪府立大学大学院工学研究科 電気・情報系専攻
知能情報工学分野
市橋秀友
TEL 072-254-9352
E-mail：ichi@cs.osakafu-u.ac.jp

著者略歴

市橋　秀友　(いちはし　ひでとも) [正会員]
1971 年大阪府立大学工学部経営工学科卒業．同年松下電器産業（株）入社，1981 年大阪府立大学工学部経営工学科助手，1987 年同講師，1989 年同助教授，1993 年同教授，2000 年同大学院工学研究科教授，現在，知能情報工学分野に所属．工学博士．ファジィクラスタリングに基づく識別器とその知能情報システムへの応用研究に従事．IEEE，日本知能情報ファジィ学会の会員．

長浦　一哉　(ながうら　かずや) [非会員]
2009 年大阪府立大学工学部知能情報工学科卒業，同大学院工学研究科知能情報工学分野在学中，ファジィクラスタリングに基づく識別器とその知能情報システムへの応用研究に従事

野津　亮　(のつ　あきら) [非会員]
2005 年 3 月京都大学大学院情報学研究科システム科学専攻博士後期課程修了．同年 4 月より大阪府立大学大学院工学研究科助手，2007 年同助教，現在に至る．博士 (情報学)．認知モデル，マルチエージェントシステムの研究に従事．計測自動制御学会，ヒューマンインタフェース学会の会員．

本多　克宏　(ほんだ　かつひろ) [正会員]
1999 年大阪府立大学大学院工学研究科博士前期課程電気・情報系専攻修了．同年日本電信電話（株）入社，同年大阪府立大学工学部経営工学科助手，2000 年同大学院工学研究科助手，2007 年同助教，2009 年同准教授，現在に至る．博士（工学）．ファジィクラスタリングによるデータマイニングやニューラルネットワークの研究に従事．IEEE，日本知能情報ファジィ学会，システム制御情報学会，日本経営工学会の会員．

Benchmarking Parameterized Fuzzy c-Means Classifier
by

Hidetomo ICHIHASHI, Kazuya NAGAURA, Akira NOTSU and Katsuhiro HONDA

Abstract: This paper reports on the performance of the fuzzy c-means based classifier (FCMC) adopting the length of cluster centers and mixing proportions of clusters as the free parameters. The FCMC consists of two phases. The first phase is an unsupervised clustering. The clustering is done on a per class basis and is implemented by using the data from one class at a time. The second phase of FCMC is a supervised classification where the free parameters of the classifier are chosen by particle swarm optimization (PSO).
High performance classifiers usually have parameters to be selected. For example, the support vector machine (SVM) has the regularization and kernel parameters. These hyperparameters are chosen by an optimization procedure to improve the generalization ability of the classifiers in terms of cross validation test. The grid search is the popular approach for SVM. Since the FCM classifier has many hyperparameters and the validation set error rate is not a unimodal function of the parameters, for the parameter search, we apply PSO inspired by social behavior of bird flocking or fish schooling. PSO is based on a simple random search and easy-to-implement. UCI benchmark datasets are used to evaluate the performance. FCM classifier in combination with the standard 10-CV procedure for parameter selection achieves better test set performance compared to k-nearest neighbor classifier. The remarkable finding is that the resubstitution (i.e., 1-CV) procedure for parameter selection also shows good test set performance. Randomized test sets performance of the classifier is comparable to that of the support vector machine (SVM) reported in the literature.
Keywords: Fuzzy c-means clustering, Classifier, Particle swarm optimization

Contact Address: Hidetomo ICHIHASHI
Department of Computer Science and Intelligent Systems, Osaka Prefecture University
1-1 Gakuen-cho, Naka, Sakai, Osaka, 599-8531, Japan
TEL : 072-254-9352
E-mail : ichi@cs.osakafu-u.ac.jp